JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

DOI: 10.1111/jedm.12419

Journal of Educational Measurement
Winter 2024, Vol. 61, No. 4, pp. 682–709

Differences in Time Usage as a Competing Hypothesis for Observed Group Differences in Accuracy with an Application to Observed Gender Differences in PISA Data

Radhika Kapoor and Erin Fahle
Stanford University
Klint Kanopka
New York University
David Klinowski
University of Pittsburgh
Ana Trindade Ribeiro and Benjamin W. Domingue
Stanford University

Group differences in test scores are a key metric in education policy. Response time offers novel opportunities for understanding these differences, especially in low-stakes settings. Here, we describe how observed group differences in test accuracy can be attributed to group differences in latent response speed or group differences in latent capacity, where capacity is defined as expected accuracy for a given response speed. This article introduces a method for decomposing observed group differences in accuracy into these differences in speed versus differences in capacity. We first illustrate in simulation studies that this approach can reliably distinguish between group speed and capacity differences. We then use this approach to probe gender differences in science and reading fluency in PISA 2018 for 71 countries. In science, score differentials largely increase when males, who respond more rapidly, are the higher performing group and decrease when females, who respond more slowly, are the higher performing group. In reading fluency, score differentials decrease where females, who respond more rapidly, are the higher performing group. This method can be used to analyze group differences especially in low-stakes assessments where there are potential group differences in speed.

Introduction

Differences in performance across groups are of persistent interest in education policy dating back to at least the 1960s (e.g., Coleman, 1968; Darling-Hammond, 2015). Such differences are often identified as group differences in test scores. These differences in outcomes across various groupings—race (Reardon, Robinson-Cimpian, and Weathers, 2014), gender (Reardon et al., 2019), or urban-rural settings (Lounkaew, 2013)—may offer information about where policy interventions are needed. However, observed group differences can be driven by a range of factors. Especially in low-stakes measures, observed differences in scores cannot be solely attributed to differences in ability; there may be other systematic differences between the groups, such as motivation (Soland, 2018a, 2018b) or use of guessing strategies (Baldiga, 2014). Further, differences in group performance can be a result

of differences in test and item characteristics, for example, item type such as multiple choice versus open-response items (Solheim and Lundetræ, 2020; Reardon et al., 2018; Shear, 2023), respondent familiarity with chosen item type (Caldwell and Pate, 2013), or other characteristics of how a construct is tested, such as the type of text used to measure reading literacy (Solheim and Lundetræ, 2018; Yousefpoori-Naeim, Bulut, and Tan, 2023), or paper versus digital medium of delivery (Cho et al., 2024). Such differences can be captured by systematic group differences in the use of time (Ercikan, Guo, and He, 2020; Guo and Ercikan, 2020). In this article, we describe an approach that uses test speed—by which we mean responses produced more rapidly as observed based on response time—as a means of testing alternative mechanisms for inducing observed differences in group-level accuracy.

As a motivating example, consider gender differences in the OECD's Programme for International Student Assessment (PISA). PISA has consistently reported that females outperform males in reading (Schleicher, 2019). Gender differences in PISA reading scores have been interpreted as indicative of differences in reading abilities and have prompted calls for curricular reform (Brozo et al., 2014). However, there are also gender differences in test-taking speed and time taken to respond to items (Schleicher, 2019); these speed differences could be due to gender differences in time use strategies, motivation, effort, and sustained performance on a test (Gneezy et al., 2019; Balart and Oosterveen, 2019; Borgovoni and Biecek, 2016; Wise and DeMars, 2006), competitiveness (Niederle and Vesterlund, 2010), exam stress (Cai et al., 2019), risk preferences (Linn et al., 1987), or anxiety (Hyde, Fennema, and Lamon, 1990). Task characteristics such as use of multiple-choice items or open response items can also induce gender differences in scores (Solheim and Lundetræ, 2020; Reardon et al., 2018). There are thus many plausible reasons we may observe male-female differences that are unrelated to ability as it is conventionally understood.

Our article uses response time recorded during the test-taking process to focus on a particular question: Could differences in test-taking speed be driving the observed group difference? For clarity of discussion, we will use the term *capacity* to indicate accuracy for a given response speed; specifically, for responses produced in the same length of time, an individual with higher capacity would be expected to produce a more accurate response. We study capacities via the conditional accuracy function (CAF) that describes the intra-individual relationship between response speed and accuracy. After identifying CAFs, we then consider competing rationales for how observed differences in accuracy may come to be. Are observed differences in accuracy more consistent with group differences in *capacity* (i.e., the groups use similar amounts of time but one group has higher capacity) or with group differences in *speed* (i.e., one group responds more rapidly than the other which has consequences for observed accuracy)? Performance differences may, of course, be due to a mixture of these mechanisms. Our aim is to illustrate such possibilities and to offer a methodology for estimating the size of such differences. Note also that we avoided invoking the conventional idea of "ability"; ability is presumably related to capacity and speed, but these relationships may be complex. We further discuss these connections in the discussion.

We proceed as follows. We first describe our approach to studying the role of test speed in explaining observed group differences. We then perform a simulation study demonstrating that the approach can parse the role of time and capacity in

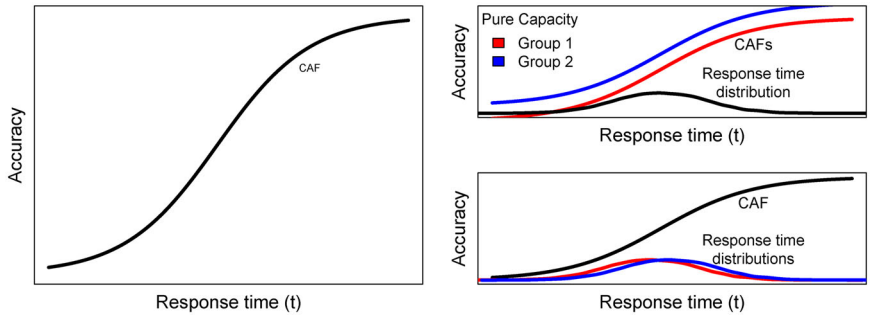

Figure 1. A conceptual overview of our approach. *X*-axis of the curves represents response time, and *Y*-axis represents predicted accuracy.

observed group differences in accuracy when, crucially, we are blind to certain components of the data generation process. We then consider gender differences in two different types of tasks from PISA 2018, scientific literacy and reading fluency, as illustrations of how our approach can be used to further interrogate the meaning of observed group differences. Scientific literacy measures knowledge of science phenomena, evaluating scientific inquiry, and interpreting data scientifically (OECD, 2019). Spending more time reflecting on the task could be a result of more cognitive processing, and successful respondents would be expected to spend more time on task or be less speedy. In contrast, reading fluency measures respondent ability to read text automatically and to "comprehend the overall meaning of the text" (OECD, 2019a). Efficient readers are hence expected to spend less time on reading fluency items. We close with a discussion of how this approach can be further used to improve resulting inferences, as well as how these group differences are understood and discussed in education policy and classroom contexts.

Time-Based Adjustments to Differences in Accuracy

We offer a conceptual presentation of our approach in Figure 1. On the left, we show a hypothetical conditional accuracy function (CAF) for an item. The CAF captures the level of accuracy for an individual were they to answer a question at different speeds (Heitz and Engle, 2007). We emphasize that it is describing a within-person phenomenon probing the degree to which a person's speed is associated with accuracy. Historically, it has been scrutinized in, for example, studies of the speed-accuracy tradeoff (Heitz, 2014) which suggests that slower responses (i.e., responses that are less hurried) are more accurate ones. These studies typically use experimental interventions designed to force respondents to work at different speeds.

More recent work focuses on identification of the CAF in observational settings (Bolsinova et al., 2017; Domingue et al., 2021, 2022; Kang, De Boeck, and Ratcliff, 2022) of the kind relevant in educational measurement. This work asks whether idiosyncratic within-person variation in speed is associated with variation in accuracy. We build on this approach here. In particular, we make use of the fact that integrating the CAF with respect to the density of response times provides us with the marginal accuracy (i.e., the proportion of correct responses).

Consider the right panels of Figure 1, which depict two hypothetical scenarios wherein we are studying performance differences between groups. The CAF is upward sloping in both scenarios, showing expected accuracy increase with increasing response time. In both panels, the blue group has a higher level of overall accuracy on a given cognitive item. In the top panel, both groups have the same distribution of response times (the black density), but each group has a distinct CAF. Respondents in the blue group are uniformly more likely to respond correctly for any given time; that is, this group has higher capacity in the sense introduced earlier. Observed differences in accuracy are due to this capacity difference (i.e., the vertical offset between the two CAFs) rather than due to a difference in speed between groups. The capacity difference suggests that for a given amount of time spent responding to an item, a blue respondent is more likely to produce a correct response than a red respondent. In the bottom panel, we offer an alternative scenario wherein the two groups have different response time distributions but the same CAF. Here, differences in observed accuracy are due to a difference in speed. The response time distribution for the blue group is to the right of the red group; that is, a student in the blue group will tend to use more time (i.e., be slower to respond or be less speedy). Given the monotonically increasing CAF, this translates into higher accuracy for the blue group.

We make several additional points about this scenario. First, while here we focus on the upward sloping CAF suggested by the speed-accuracy tradeoff, our approach is flexible and does not presuppose the shape of the CAF. For example, van der Linden's hierarchical model (Van der Linden, 2007) builds upon the assumption of a flat CAF such that no changes in accuracy result for within-person variation in time usage. Empirically, a range of relationships seem possible (Domingue et al., 2022; Chen et al., 2018; Mutak et al., 2024), including downward-sloping CAFs or curvilinear CAFs where accuracy increases when more time is spent (consistent with the speed-accuracy tradeoff) but only up to a point and then decreases (possibly due to disengagement or confusion). Second, conventional ability estimates that are based purely on observed accuracy would be unable to distinguish between the speed- and capacity-driven differences in accuracy that we discuss above; identification of these differences obviously requires observation of time as well. Given that time usage may differ across groups for reasons unrelated to the construct of interest, attempting to partition observed differences in this way thus seems desirable. Third, in contrast to the differences we show here, which are due purely to speed or capacity, empirically we may observe tangled mixtures of both speed and capacity-driven differences in accuracy, including cases where both capacity and speed are functions of ability (as conventionally understood). In these cases, latent speed and capacity together influence accuracy. The method introduced in this article does not make strong assumptions about the relationship between latent capacity and speed. These assumptions are discussed in more detail in "Simulation Study" section. Our simulation below helps to clarify how our approach may work in such a scenario. We now discuss the methodological details of this approach.

Identification of the CAF

In Equation 1, we use a linear probability model to model responses $y_{ij} \in \{0, 1\}$, where y_{ij} is a response and t_{ij} is the time required to respond from student i to item

j, $\boldsymbol{b}(\log t_{ij})$ represents b-splines[1] of response time t_{ij}, λ_i and γ_j are person and item fixed effects, and $\epsilon \sim N(0, \sigma_y^2)$:

$$y_{ij} = \boldsymbol{\beta} \cdot \boldsymbol{b}(\log t_{ij}) + \lambda_i + \gamma_j + \epsilon \tag{1}$$

$$h(t_{ij}) = \boldsymbol{\beta} \cdot \boldsymbol{b}(\log t_{ij}). \tag{2}$$

Note that the term $\boldsymbol{b}(\log t_{ij})$ describes the expected accuracy as a function of t_{ij}; that is, it is the CAF. In Equation 2, we denote this term as $h(t_{ij})$. Given that the form of h is unknown, we consider a flexible, spline-based approach (Domingue et al., 2021, 2022) for identification of h. Note that, as in other studies (Van der Linden, 2006; Chen et al., 2018), we analyze the log of response time.

We make two additional notes about this approach. First, below we will consider CAFs estimated separately across groups. Doing so will merely require analysis of responses y_{ij} solely for individual i in a specific group. Second, we rely upon a linear probability model for computational reasons. In particular, it allows for inclusion of large number of fixed effects, as our goal is understanding within-person tradeoffs between speed and accuracy; in simulations (Domingue et al., 2022, 2021), this approach works well in a variety of settings and data generation conditions. This specification assumes that the functional forms of CAFs are similar across items, so it is most appropriate in cases where items are relatively homogeneous.

CAFs and Observed Accuracy

Having described the CAF (i.e., h) in Equations 1 and 2, we can now use it to decompose accuracy differences between groups. Given that we study group differences, we emphasize that individual respondents i will be denoted as being a member of group $g \in \{1, 2\}$. As we are concerned with group-level differences in accuracy, we denote these as A where $A = A_2 - A_1$ and

$$A_g = \frac{1}{J \cdot N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{J} y_{ij}, \tag{3}$$

where N_g is the number of respondents in Group g where $g \in \{1, 2\}$ and y_{ij} is the response by person i to item j. Note that the difference A will contain all speed- and capacity-based differences. Our attempt to untangle speed- and capacity-based differences will be built on a consideration of group differences in both CAFs and response time distributions.

We now turn to an approach for estimating observed accuracy that relies on CAFs; this approach is designed to help partition differences into those driven by time usage and those driven by capacity. First, we can estimate the CAF for group 1, $h_1(t)$, using Equation 2. Next, Group 1 mean accuracy ($\widehat{A_1}$) can be estimated by integrating this CAF $h_1(t)$ over the density of the response time distribution for Group 1

$$\widehat{A_1} = \int_{T_1} h_1(t_1)\, dt_1. \tag{4}$$

Here, T_1 represents response time distribution of Group 1 and dt_1 denotes the response time density $f(t_1)dt_1$. We can extend this idea to also estimate the group difference $\widehat{A}$

$$\widehat{A} = \int_{T_2} h_2\, dt_2 - \int_{T_1} h_1\, dt_1 = \widehat{A}_2 - \widehat{A}_1, \tag{5}$$

where $\widehat{A_2}$ is estimated by integrating $h_2(t)$, the CAF for group 2, over the group response time distribution T_2. Note that $\widehat{A}$ will continue to contain information about both speed and capacity differences. We now turn to alternative indices that will allow us to untangle the roles of speed and capacity.

We first study the performance difference due purely to capacity as in the top right panel of Figure 1. We want a quantity that should index the difference in CAF shown in that panel; that is, we want an index that conveys information about the observed difference in performance *due purely to capacity*. We compute

$$\widehat{A}_c = \int_T (h_2 - h_1)\, dt = \int_T h_2\, dt - \int_T h_1\, dt, \tag{6}$$

where we are now integrating with respect to the overall time distribution across both groups (T). $\widehat{A_c}$ is informative about the difference in accuracy due to differences in capacity, assuming common time usage. Empirically, we analyze demeaned response time ($\log t_{ij} - \frac{1}{N}\sum_{i=1}^{N} \log t_{ij}$ for each item j and N respondents) to improve interpretability of the results.

We next compute group differences due to speed while assuming common capacity as described by the CAF (i.e., the bottom right of Figure 1)

$$\widehat{A}_s = \int_{T_2} h\, dt_2 - \int_{T_1} h\, dt_1. \tag{7}$$

Here, h is an overall CAF (as in the bottom-right of Figure 1) estimated using data from both groups. We then integrate it with respect to each group's response time distribution. $\widehat{A_s}$ is then informative about the difference in accuracy due to different usage of time while assuming a common CAF. Note that this approach decomposes group differences $\widehat{A_2} - \widehat{A_1}$, denoted by $\widehat{A_{est}}$ into differences due to speed and differences due to capacity[2]

$$\begin{aligned}
\widehat{A}_c + \widehat{A}_s &= \left(\int_T h_2\, dt - \int_T h_1\, dt\right) + \left(\int_{T_2} h\, dt_2 - \int_{T_1} h\, dt_1\right) \\
&\approx \left(\int_{T_2} h_2\, dt_2 - \int_{T_2} h_1\, dt_2\right) + \left(\int_{T_2} h_1\, dt_2 - \int_{T_1} h_1\, dt_1\right) \\
&= \int_{T_2} h_2\, dt_2 - \int_{T_1} h_1\, dt_1 = \widehat{A}_{\text{est}}.
\end{aligned} \tag{8}$$

Simulation Study

Data Generation

We describe a simulation study focused on using $\widehat{A_c}$ and $\widehat{A_s}$ to capture known differences in speed and capacity under varying assumptions; in particular, we test

if we can separate group difference attributable to speed and to capacity using these quantities. For a given number of respondents, N, we simulate item response data for a 45 item test as follows:

- Individual capacity (θ_i) and speed parameters (τ_i) were drawn for the $N_1 (= N/2)$ respondents in group 1 from a bivariate normal distribution with mean zero, unit standard deviations, and covariance ρ.
- Individual capacity and speed parameters for the $N_2 (= N/2)$ respondents in group 2 were drawn from a bivariate normal distribution with mean $\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}$ but the same variance-covariance matrix as in group 1 (i.e., $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$).
- We draw 45 difficulty parameters (δ_j) from a Normal distribution

$$\delta_j \sim \text{Normal}\big(0, .5^2\big). \tag{9}$$

- Response times for an individual response are generated as

$$\log t_{ij} \sim \text{Normal}\big(.1 + .3(-1\tau_i), .5^2\big), \tag{10}$$

  where the coefficient on τ_i is negative so that τ_i can be interpreted as a person's latent speed. This leads to $\mathbb{E}(t_{ij}) = \exp(.1 + \frac{.5^2}{2}) = 1.3$ (IQR $= 1.9s$).[3] Note that all items are assumed to have the same expected response time (i.e., there is no variation in the time intensity of the items).
- Building on other work (Wang and Hanson, 2005), we simulate item responses via Equation 11:

$$y_{ij} \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(-1(\theta_i + b_t t_{ij} - \delta_j))}\right). \tag{11}$$

  The parameter b_t controls the effect of the response time for an individual item response on accuracy and the shape of CAF; if $b_t > 0$ then the CAF is upward sloping (as in the left panel of Figure 1) such that more time spent generating a response for an individual should predict higher levels of accuracy.

Data generation requires specification of parameters (N, μ_θ, μ_τ, ρ, b_t). Here, we vary $b_t \in \{.5, 0, -.5\}$ and fix $\rho = .25$. We choose $N = 1,000$ to ensure that the approach works with relatively modestly sized samples. Appendix 2 elaborates on results for $N = 10,000$, and other values for number of items and latent capacity-speed covariance ρ. Results were not sensitive to choice of parameters.[4] Code to replicate our simulation results is available.[5]

We illustrate the effect of Equation 11 on the relevant item response function (IRF) in the left panel of Figure 2. Under this model, responses that take more time (emphasized as redder curves) are more likely to be accurate for a given level of θ. The right panel of Figure 2 considers the CAFs for different values of θ. Note that they are all upward sloping, a consequence of the fact that $b_t > 0$. CAFs need not have this shape; some models—specifically the hierarchical model for response time and accuracy (Van der Linden, 2007)—posit no within-person interplay between speed and accuracy; under the model, the resulting CAFs would be flat (i.e., within-person changes in speed are unassociated with accuracy).

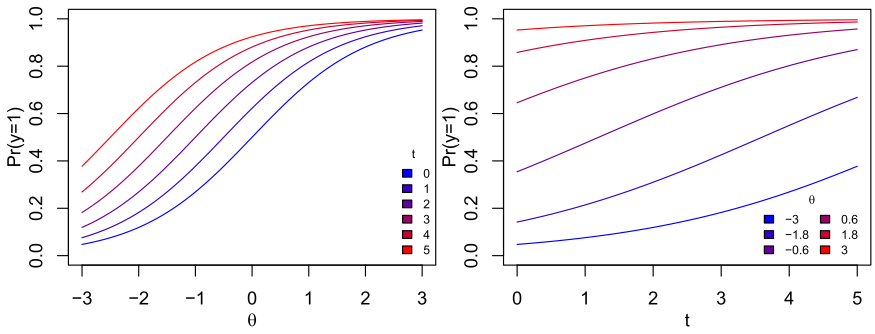

Figure 2. Left: IRFs for different response times (t) based on Equation 11. *Right*: CAFs for different values of θ. We set $\delta_j = 0$ and $b_t = .5$ throughout.

Our approach is flexible and able to identify a range of potential CAFs (Domingue et al., 2022). Note also one implication of the right panel of Figure 2. Variation of μ_θ will have the effect of inducing variation in the "level" of the CAF along the lines of the capacity-based difference in the top right panel of Figure 1. In this sense, capacity differences are due to μ_θ. However, we have refrained from calling them ability-based differences here as conventional inferences about observed group differences have tended to assert that they are "ability" differences. Our point is that such differences can be due to what we call capacity (i.e., μ_θ) or speed (i.e., μ_τ), and we believe the novel terminology helps to emphasize this difference.

Given their relationship with capacity and speed, the μ_θ and μ_τ parameters are critical. As noted above, μ_θ will effectively index capacity-based differences; we choose 100 values for μ_θ from Unif(0, .25). Speed-based variation in observed accuracy is controlled via μ_τ. Note that we can choose one group to have the larger capacity (which we do here given that $\mu_\theta > 0$) but still allow for the full range of possibilities by allowing μ_τ to be positive, zero, or negative; we thus choose $\mu_\tau \in \{-.5, 0, .5\}$. Variation of both these parameters allows observed accuracy differences to be due to either speed or capacity; the key question in simulation studies is whether the $\widehat{A_s}$ and $\widehat{A_c}$ estimates can reveal what is known to be happening in these simulations.

Results of Simulation Study

We first consider estimates of accuracy due to variation of key simulation parameters $\{\mu_\tau, \mu_\theta, b_t\}$ in Figure 3. Dot color represents the value of μ_τ throughout, where slower respondents ($\mu_\tau = -.5$) are in red, the faster respondents are in blue ($\mu_\tau = .5$). Figure 3 shows the observed differences in accuracy (A) as a function of μ_θ and b_t. The observed differences vary widely but note that they are stratified by μ_τ (i.e., similar colors tend to group together along diagonal lines). In the left panel, slower respondents ($\mu_\tau = -.5$) are more accurate ($b_t > 0$); this is reflected in the red dots generally having the largest A values. This direction is reversed when CAFs are downward sloping ($b_t = -.5$). In the right panel, note that the blue dots representing the group with more speed are now on top. In practice, we would not observe μ_τ and so may incorrectly attribute variation in A entirely to variation in capacity. We now

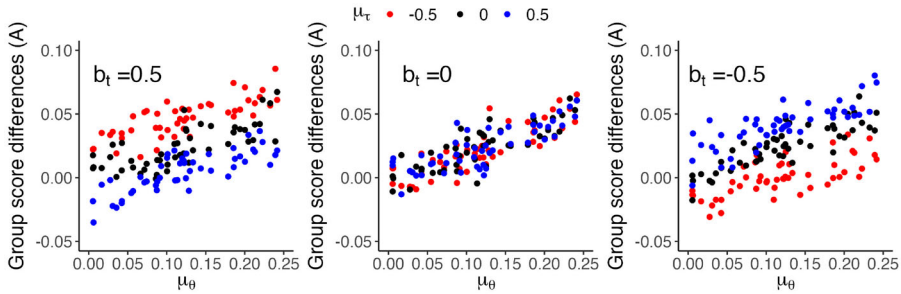

Figure 3. Results from simulation study for 100 values of μ_θ for each set of simulation conditions (N, μ_τ, b_t) Here rows represent N=1,000, columns represent $b_t \in \{.5, 0, -.5\}$, μ_τ is represented by the color of the points. Here we focus on estimates of observed differences, A, as a function of μ_θ. Note that a given μ_θ induces a wide range of A values and these values are stratified by the level of μ_τ; it is this stratification that we aim to eliminate. *Left*: When the CAF is upward sloping ($b_t = .5$), the slower group (red points; $\mu_\tau = -.5$) would be observed to have a higher score, for the same capacity (μ_θ) level. *Middle*: When CAFs are flat ($b_t = 0$), we do not observe a score differential for the same μ_θ. *Right*: The original order is reversed when CAFs are downward sloping ($b_t = -.5$); that is, the slower group has the lower score at the same capacity level.

turn to the performance of the $\widehat{A_c}$ indices in decomposing observed differences in A to group differences in capacity.

In Figure 4, we first consider $\widehat{A_c}$ as a function of μ_θ with color continuing to index μ_τ (top row). We see that $\widehat{A_c}$ collapses the previously stratified differences across μ_τ such that the colored points are effectively overlapping; remaining differences in $\widehat{A_c}$ are due purely to differences in μ_θ (note the clear association between μ_θ and $\widehat{A_c}$). This is due to the fact that $\widehat{A_c}$ removes the effect of speed—that is, color in Figure 4—on the dots' placement. When $\mu_\tau \neq 0$ there is a clear variation in the average A value as a function of μ_τ but, as desired, no variation in $\widehat{A_c}$.

The difference between $\widehat{A_c}$ alongside the naive estimate A for $N = 1,000$ is emphasized in the bottom row of Figure 4. In the bottom left panel, we note that the blue dots lie above the 45° line; that is, adjusted differences are larger than observed differences when the higher capacity group is faster (more speedy) and CAFs are upward sloping. However, if the higher capacity group was already spending more time on items and CAFs were upward sloping, adjusted differences would decrease. This corresponds to the red dots in the bottom left panel. Note that these predicted differences flip if CAFs were downward sloping, shown in the bottom right panel. The blue dots here represent the higher capacity group which is more speedy. Adjusted differences decrease in this case, as the CAF is downward sloping and there are negative returns to spending more time on items.

The recovery of group difference due to speed, $\widehat{A_s}$, is discussed in Appendix 1. Specifically, we suspect that interest is typically in the $\widehat{A_c}$ quantities that focus on group differences in μ_θ after removing the effect of speed differences. Simulation results were not sensitive to choice of number of respondents, number of items, and latent capacity-speed covariance ρ. These results are discussed in Appendix 2. Note

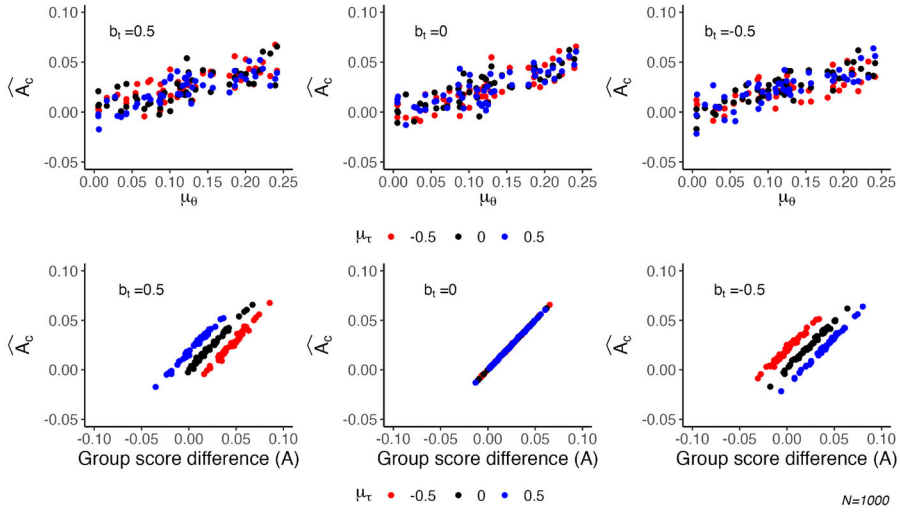

Figure 4. Results from simulation study for 100 values of μ_θ for each set of simulation conditions (N, μ_τ), when $N = 1,000$. *Top* shows the association between μ_θ and the adjusted score $\widehat{A}_c$, for $b_t \in \{.5, 0, -.5\}$. Note that after adjusting the score for CAF differences, the score difference $\widehat{A}_c$ is still highly associated with μ_θ, but not with μ_τ, for all values of b_t. The *bottom row* column shows the association between observed difference A and $\widehat{A}_c$. Note that with $\widehat{A}_c$, adjusted scores are higher for the slower blue group when CAFs are upward sloping ($b_t > 0$ in *Left*), and lower when CAFs are downward sloping ($b_t < 0$ in *Right*)

that this setup also assumes that individuals are attempting the same set of items, that groups have similar CAFs, and groups are relatively well-balanced. If groups are not well balanced, the behavior of the aggregate CAF (h in Equation 7) and aggregate time density (dt in Equation 6) would be overly influenced by one group; we comment further on this matter in the discussion. Given that this simulation study shows that we can reliably detect speed- and capacity-based differences when these assumptions are met, we move to an empirical analysis focusing on these quantities.

Empirical Analysis

PISA Data

We use publicly available data from PISA 2018[6] scientific literacy and reading fluency assessments from 71 countries to showcase the approach described here. Data processing and analysis is conducted separately for each country.[7] Table 1 gives an overview of the data. As expected, roughly half the respondents were female (corresponding to the simulation where group sizes N_1 and N_2 were equal sized). Response time for an item is defined as the total time respondent spent across multiple "visits" by a respondent.[8] To minimize noise due to, for example, rapid random guessing or responses that involved technology issues, we exclude responses from the first and hundredth percentiles. Given the skew in time, we analyze the log of time (in seconds). For the analysis, log response time was centered at zero, following Chen et al. (2018). We focus interpretation on $\widehat{A}_c$ values.

Table 1
Descriptive Summary

	Science PISA 2018	Reading Fluency PISA 2018
N items	83	65
N respondents	109,506	545,756
N countries	71	71
% Female	50%	50%
Accuracy	.49	.88
Mean score difference (female-male)	.00%	2.34%
Median score difference (female-male)	−.07%	2.12%
75th quantile score difference (female-male)	−1.31%	1.50%
25th quantile score difference (female-male)	.88%	2.77%
Median time (seconds)	49.8	3.6
Mean Kullback-Leibler (KL) distance	.026	.034

Note that accuracy and response time behaviors are different for reading fluency and scientific literacy items. Reading fluency items were relatively easy (average accuracy of 88%) while scientific literacy items are much more difficult (average accuracy of 49%). Females outperformed males on reading fluency items in all countries while the direction of the performance gap was more mixed across countries for scientific literacy. As we may anticipate, respondents spent more time on scientific literacy items; respondents spent median response time of 49.8 seconds on the scientific literacy items as compared to 3.6 seconds on reading fluency items. However, gender differences in response time distributions were slightly bigger in reading fluency when compared to science, as measured by the KL distance between male and female response time distributions.[9] To clarify the key ideas from this analysis, we focus on results from four countries for illustration: Brazil (BRA), Netherlands (NLD), Moscow Region of Russia (QMR), and the United Arab Emirates (ARE).[10]

Empirical Study 1: PISA 2018 Science Items

Data. This section uses publicly available data from the PISA 2018 Science assessment. Items in this domain assess if students can interpret and make inferences about scientific phenomenon, while focusing on a range of science competencies (OECD, 2019). Correct responses to science items require understanding of items, recall of relevant scientific knowledge, as well as higher order cognitive reasoning.[11] As respondents would need to spend time on reading and reasoning about the problem, we anticipate that more careful responses (i.e., slower responses) will generally be more accurate.

Figure 5 describes scientific literacy response and time data from 71 countries. To avoid item order effects, we include dichotomous, multiple-choice items from Booklet 13 to 18.[12] We discuss the panels in the figure from left to right. *Panel 1* shows the density of observed score differences between female and male groups across countries. Note that the density distribution is centered around zero and is

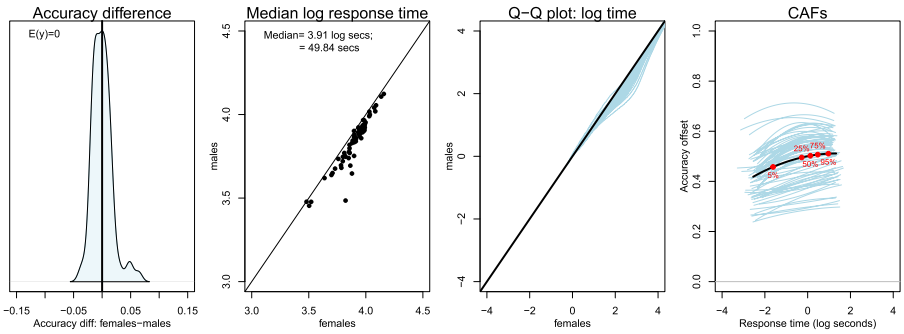

Figure 5. Gender difference in time usage, score, and CAF shape for PISA Scientific literacy. From left to right—*Panel 1. Accuracy difference*: Density distribution of difference of female and male observed accuracy scores (averaged across items at the country level). Vertical line is added at x=0. *Panel 2. Median log response time*: Each point represents a country. *Y*-axis represents median log response time for males and *X*-axis for females. *Panel 3. Q-Q plots*: Quantile-quantile (Q-Q) plots of male and female log response time. *Panel 4. CAFs*: Blue lines show country-specific CAFs, mapped for 1st-99th percentile of each gender's centered log response time distribution. Overall CAF is shown in black along with indications of where certain percentiles of the complete response time distribution occur.

roughly symmetric, that is, countries in which males outperform females are roughly equal in number to the countries where females outperform males.

Panel 2 plots the median response time for males (*Y*-axis) and female gender (*X*-axis), where each point represents a country. The points largely lie below the 45° line, showing that on average, males are spending less time on science items when compared to females. *Panel 3* shows quantile-quantile (Q-Q) plots of response time distributions for female and male groups. Each blue line in the plot represents the Q-Q plot for one country. A blue line lies along the diagonal when male and female response times have similar distributions. In *Panel 3*, blue lines lie along the diagonal for the lower and middle ends of the response time distributions, and dip below the diagonal on the right end for some countries. This dip represents a relative right skew for male response time distributions compared to females. *Panel 4* shows CAFs for all countries in light blue, with an overall CAF shown in black for emphasis. CAFs are generally upward sloping implying that less rapid responses are more accurate. However, as response time increases, there are reductions in gains associated with increases in marginal time.[13]

Country-level adjusted differences. Figure 6 shows the adjusted scores $\widehat{A_c}$ mapped against observed scores A. The mean change in scores $(A - A_c)$ across countries is .15% (SD = .35%). To avoid confusion, results are split by values of A such that the left panel focuses on countries where males are the higher performance group and the right panel shows countries where females are higher performing. Country-level CAFs and response time distributions for the selected four example countries (Brazil, Netherlands, Moscow Region of Russia, and United Arab Emirates) are shown in the Appendix 4. Note that the CAFs are largely upward sloping for scientific literacy, corresponding to the left panel of Figures 3 and 4.

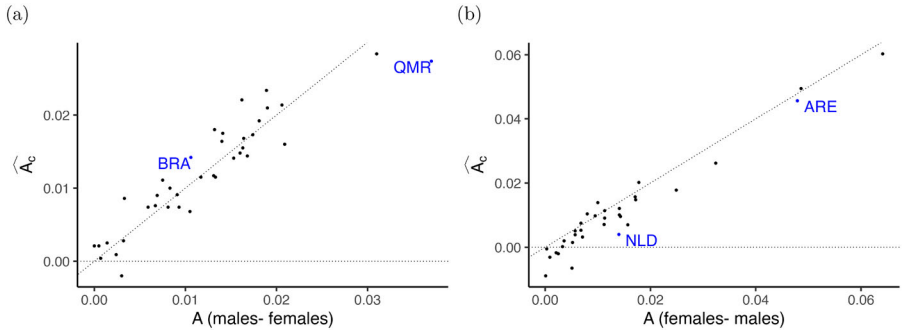

Figure 6. Graph shows group difference estimates corrected for speed $\widehat{A_c}$, compared to observed scores A. Observed score differences A were biased estimates of capacity. Each point represents a country. Dashed diagonal line is the 45° line, where $A = \widehat{A_c}$; horizontal line is at $\widehat{A_c} = 0$. Selected countries for deep dive are highlighted in blue. *Left*: Graph shows results for countries where males had scored more than females, where adjusted score largely increases. *Right*: Graph shows results for countries where females had scored more than males, where adjusted scores largely decreased.

If the group with faster respondents was also performing worse, we would expect observed scores to overestimate group capacity differences; that is, group scores after adjusting for speed differences would reduce ($\widehat{A_c} < A$). However, note that group differences in accuracy and speed are relatively small. We discuss in turn the range of possible scenarios.

Turning to the $\widehat{A_c}$ values (results related to $\widehat{A_s}$ are reported in the Appendix 1), we observe that adjusted scores increase for most countries when males are the higher performing group (Figure 6a). Consider Brazil: males are the higher performing group but also respond more rapidly (i.e., this scenario is similar to the blue group in the left panel of Figure 4). Adjusting for time usage differences should thus increase $\widehat{A_c}$ relative to $\widehat{A}$ the difference, corresponding to the bottom left panel in Figure 4. For Brazil, relative to an observed group score difference of $A = 1.06\%$, we estimate $\widehat{A_c} = 1.42\%$. From this perspective, the observed score difference is a potential underestimate of the difference in capacity. The second observation in the left panel is Moscow, where observed score differences decrease from 3.7% to 2.7%. Males are again the higher performing group but they respond more slowly; that is, they spend more time on the item. This corresponds to the red group in the lower left panel of Figure 4.

In the right panel (Figure 6b), we observe that adjusted scores decrease for most countries when females are the higher performing group. We focus on results for two countries: Netherlands (NLD) and United Arab Emirates (ARE). In both countries, females are the higher performing group that also spend more time on the item. This again corresponds to the bottom left panel of Figure 4, where females correspond to the red dots. Hence, observed group differences are expected to decrease. In Netherlands, group differences decrease from 1.4% to almost 0%; in United Arab Emirates, differences decrease from 4.8% to 4.6%. Details of country-level results for these four countries are discussed in Appendix 4.

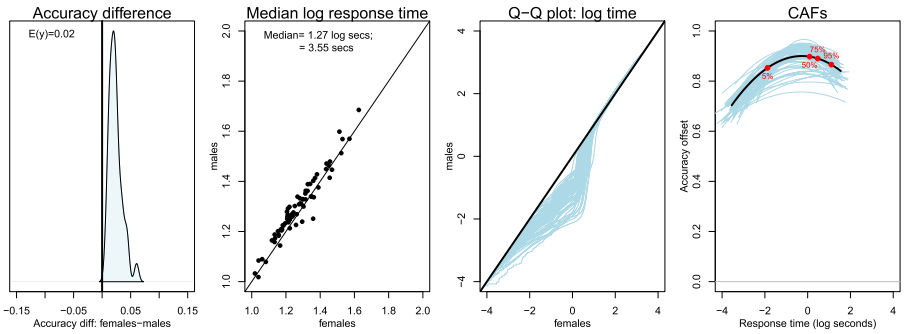

Figure 7. Gender difference in time usage, score, and CAF shape. Vertical line is plotted at $x = 0$. From left to right—*Panel 1. Accuracy difference*: Density distribution of difference of female and male observed accuracy scores (averaged across items at the country level). Vertical line is added at $x = 0$. *Panel 2. Median log response time*: Each point represents a country. *Y*-axis represents median log response time for males and *X*-axis for females. *Panel 3. Q-Q plots*: Quantile-quantile (Q-Q) plots of male and female logged response time. *Panel 4. CAFs*: Blue lines show country-specific CAFs, mapped for 1st-99th percentile of each gender's centered log response time distribution. Overall CAF is shown in black along with indications of where certain percentiles of the complete response time distribution occur.

In both panels, we also observe some countries lie along the 45° line, where observed and adjusted differences are very close. In these cases, observed group score differences can be attributed to group differences in capacity; that is, there is no observable impact of adjusting for group differences in speed. In the next section, we discuss the results for PISA reading fluency items.

Empirical Study 2: PISA 2018 Reading Fluency

Data. We now discuss analysis of the PISA 2018 for Reading Fluency.[14] The PISA 2018 reading framework organized the cognitive processes involved in reading as: (1) reading fluently and (2) locating information, understanding, and evaluating and reflecting (OECD, 2019a). We focus on reading items written to target the first process; we will henceforth refer to these as Reading Fluency or Fluency items.

Fluency items in PISA 2018 were designed to evaluate if respondents could read and understand simple text efficiently. Students were presented 21-22 sentences in 3 minutes, and they had to determine if the sentence made sense or not (items had two possible responses: yes or no). An example sentence is "*The window sang loudly*" (OECD, 2019b). Response time is hence small (median time is 3.6 seconds in Table 1), and faster responses might be expected to be more accurate since this task requires automaticity in reading.

In Figure 7, *Panel 1*, we observe that females were the more accurate group across countries, similar to the findings in Table 1. The percentage change in group mean score difference across countries was 2.34% (IQR 1.27%). Our findings are consistent with those from the PISA 2018 report (Schleicher, 2019), which reported higher reading scores for females for most countries.

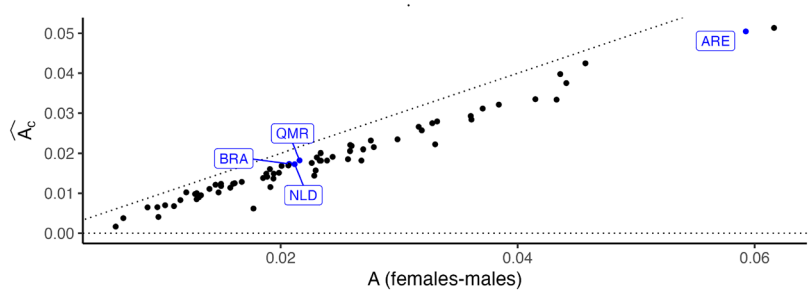

Figure 8. Adjusted scores ($\widehat{A_c}$) vs. observed scores (A), each point represents a country. Dashed diagonal line is at 45°, where $A = \widehat{A_c}$. Horizontal line is at $\widehat{A_c} = 0$.

In *Panel 2*, each point maps median response time for males in a country to the median female response time. The points lie above the 45° line for most countries, showing that females spent less time on items compared to males on average. The response time distributions also show a bigger skew and more variation in differences (see *Panel 3*, which contains Q-Q plots for male and female log response time distributions). The CAFs in *Panel 4* are non-monotonic suggesting that returns to increases in time do not always yield increases in accuracy. In the case of reading fluency, on average, the CAFs are downward sloping to the right of mean response time; note that mean response time is centered at 0. As we shall see below, this centering has important implications for interpreting results.

Country-level adjusted differences. We now consider adjustments in observed differences in accuracy for differences in time usage. Note that the CAFs have a negative slope at the mean log response time (Figure 7, right panel). In this case, it seems that generation of a correct response typically happens rapidly and additional time is perhaps indicative of wheel-spinning and other unproductive activities (Smallwood and Schooler, 2015).

We again consider the the A and $\widehat{A_c}$ values for each country (Figure 8). Since females are always higher performing we only need a single panel and the X-axis represents $A_{\text{female}} - A_{\text{male}}$. Females tend to spend less time responding to these items (note that the blue lines tend to be below the diagonal in Figure 7, middle panel), and given the shape of the CAFs, we anticipate $\widehat{A_c} < A$ (similar to blue points in the bottom right panel of Figure 4, which are below the 45° line). This is the result we observe in Figure 8. Adjusting for speed differences reduces observed mean score differences between groups across countries. Across countries, the mean size of score adjustment ($A - A_c$) is .51% (*SD* .21%). Country-level results are shown in Appendix 4 for Brazil, Netherlands, Moscow region, and UAE. Group differences reduced from 2% to 1.7% for Brazil, 2.1% to 1.7% for Netherlands, 2.2% to 1.8% for Moscow, and 5.9% to 5% for UAE.

Discussion

We offer an approach designed to assess the extent to which observed group-level differences in accuracy may be due to differences in speed. We demonstrate the util-

ity of this approach in simulations studies and then consider an empirical example using PISA data. While we consider the case of gender as the grouping variable in the PISA data, the approach is generic and could be used to study other group-level differences. Our approach offers one opportunity for further studying group differences in low-stakes assessments and also raises interesting questions about the meaning of "ability"; we discuss both of these matters below.

Our empirical work focused on data from the 2018 PISA from 71 countries, with focus on results from four countries for exposition: Brazil, Netherlands, Moscow region of Russia, and United Arab Emirates. We focused on two assessments: (1) Scientific literacy items and (2) Reading Fluency. In general, previous work suggests that males are less engaged and exert less effort in low-stakes exams like PISA. We argue that this is observed in the case of scientific literacy as a reduction in time usage. Our approach offers a means of quantifying the implications of this phenomenon and adds an alternative interpretative lens that might be important as policy makers consider these results. Our findings for the fluency items connect to other results from analysis of similar items (Ranger, Kuhn, and Pohl, 2021; Goldhammer et al., 2014; Wang and Chen, 2020) wherein spending more time on items does not result in increased accuracy and may instead reflect confusion or uncertainty.

The proposed method can be used to adjust score differences for speed without making any strong assumptions about the underlying relationship between speed and capacity. This has implications for how group differences in test scores are interpreted. When the proposed method leads to observable changes in gender score differences, there would be a need for nuanced interpretation of the test scores. Similarly, the finding that observed and adjusted score differences are similar would imply that observed score differences between groups reflect group capacity differences, and can be interpreted as such. We view this analysis as similar to what is frequently done with differential item functioning (DIF) studies: null results may help ensure clarity about the nature of the inference drawn from a given measurement exercise.

This method also offers a useful tool for interrogating group differences, especially in low-stakes settings. In doing so, we invoked the notion of "capacity" and this merits additional comment. Breaking accuracy differences down into components due to capacity versus those due to speed does not necessarily imply specifics regarding their meaning. Here, we have discussed reasons for gender differences in the speed of response—for example, stress (Cai et al., 2019)—that are potentially unrelated to the cognitive processes that the items seek to discern. However, in other cases, it may not be so obvious. For example, suppose a student's response behavior is governed by a particular CAF and a particular choice of speed. What would additional instruction related to the relevant skill (e.g., instruction in reading) change in this scenario? It may change the level of the CAF such that, for a response generated in some specific length of time, we would anticipate a more accurate response for the student following the intervention. Alternatively, this instruction may primarily change the student's preference with respect to speed. The instruction could also change both CAF level and speed, and the nature of the

change could depend on both the construct and the nature of the intervention. These are complex questions that future research may care to consider. A next step for such work is to understand how respondent, task and context characteristics result in group differences in capacity, speed, and shape and level of CAFs, and in turn, how these differences translate into observed score differences. We think that the decomposition described here may be useful for describing these differences (especially when we anticipate groups behaving differently for reasons unrelated to the construct) even if it does not resolve the conceptual meaning of the differences in all cases.

There are other limitations of the current work that merit attention. First, we have presumed that the CAFs are common across items; that is, we assume the same functional form for CAFs across items and do not introduce any item-level variation. Our approach is thus most appropriate for items that are relatively homogeneous. Future work may wish to probe the degree to which there is such variation and, when it exists, whether it can be leveraged to refine the resulting inferences. Second, and related to the notion of group-level variation in the CAF, we rely upon the fact that groups are relatively balanced (as a function of the design of PISA) in terms of sample size. If they were not, this would lead to a CAF that was overly reflective of behavior in one group. Approaches based on this method that allow for unbalanced groups would be possible; research from the literature on item bias pertaining to sample size concerns (e.g., Scott et al., 2009) may be useful in developing such extensions. Third, we do not currently quantify uncertainty in the estimation of the CAF. Fourth, future work can consider how the nature of CAF and ability estimates change as a function of interventions. Fifth, this work uses linear probability models with b-splines to flexibly estimate the relationship between response time and accuracy; while the assumptions of the linear probability model are violated in the case of dichotomous responses, we think this is defensible (see arguments in Domingue et al., 2022, 2021). However, other approaches may be considered in subsequent work. Finally, the method could potentially be extended to the study of more than two groups to examine intersectionality in group differences.

Across the globe, educational outcomes are widely used to understand educational practice, often with data collected in low-stakes assessments. Group differences are a critical metric in many settings. The rapid increase in digitally administered assessments also means that response time is increasingly available. The approach described here is aimed at using such response time data to further interrogate observed group differences in accuracy. The approach works well in simulation studies with modestly sized samples. We illustrated its use with an example focusing on gender differences on the PISA assessment. However, there is interest in many such group differences; for example, cultural and language groups may utilize different response processes (Guo and Ercikan, 2020). Such differences would have implications for how group differences in test scores are interpreted, and also for how feedback might be provided to students, teachers, and schools. Approaches like ours may help add nuance related to readily observed differences in accuracy that may be useful in the conversations spurred by such feedback.

Acknowledgments

This work was funded in part by the Jacobs Foundation and was supported by Stanford Data Science Scholars Program. We would like to thank James Soland for his helpful comments on the article.

Notes

[1]We use b-splines which are a map from $\mathbb{R}^1$ to $\mathbb{R}^J$, here we use $J = 2$ (using the `bs` function in `R`), where degree 2 is chosen to describe curvilinear CAFs; this use is similar to previous work (Domingue et al., 2021). Simulations and empirical work from (Domingue et al., 2021) suggest that results are robust to the choice of J.

[2]The first part of the decomposition in Equation 8 considers $A_c = \int_T h_2\, dt - \int_T h_1\, dt$. The key assumption here is that the difference between group CAFs when integrated over joint time density is roughly similar to the difference when integrated over group 2 time density. The joint response time density does not have to be the same as group 2 time density; the requirement is for the difference. The second term considers $A_s = \int_{T_2} h\, dt_2 - \int_{T_1} h\, dt_1$. Again, we don't require joint CAF h to be the same as CAF for group 1 h_1. We only need the group differences to be similar after CAFs are integrated over response time distributions, which is a reasonable assumption. These assumptions were met in the empirical cases we discuss in the article.

[3]This response time reflects response time for simpler tasks. Simulation results are not sensitive to the choice of mean response time, since we estimate CAFs with item fixed effects.

[4]In preliminary analyses, we verified that if $b_t = 0$ the resulting effect is that values of μ_τ are not associated with group differences in accuracy.

[5]See https://github.com/rkap786/PISAhelper

[6]https://www.oecd.org/pisa/data/2018database/

[7]We take raw response and response time for items and students. Missing responses (due to adaptive nature of the test or due to students not responding, etc.) are dropped from the analysis.

[8]In PISA, respondents can visit an item more than once; that is, they can attempt it, move to another item in the same booklet, and come back. Response time is aggregated across these visits.

[9]Computed as difference between response time density for females and males using KL-distance. Average of female-male and male-female distance for each country

[10]Supplementary Index in Appendix 3 contains descriptive analysis and results for science and reading fluency for all 71 countries

[11]Example of science items can be seen here: https://t.ly/ScienceReleasedItems

[12]Item order effects in computer adaptive testing can influence item parameter estimation (Kanopka and Domingue, 2022; Leary and Dorans, 1982). Preliminary work suggests that order effects can interfere with estimation of the CAF. This is consistent with observations in other data that time usage can have different implications as a function of item order (Domingue et al., 2022) or that order effects can have implications for accuracy (Kanopka and Domingue, 2022).

[13]Note that response times are centered at 0 in *Panel 4*.

[14]Exemplars of Fluency items are available: https://www.oecd-ilibrary.org/sites/098bab1a-en/index.html?itemId=/content/component/098bab1a-en

References

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 1–11.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448.

Bolsinova, M., Tijmstra, J., Molenaar, D., & P. De Boeck., (2027). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in psychology*, *8*, 236022.

Borgovoni, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences*, *49*, 128–137.

Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement: Lessons from five PISA countries. *Journal of Adolescent & Adult Literacy*, *57*(7), 584–593.

Cai, X., Lu, Y., Pan, J., & Zhong, S. (2019). Gender gap under pressure: evidence from China's National College entrance examination. *Review of Economics and Statistics*, *101*(2), 249–263.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4), 71.

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23.

Cho, S.-J., Goodwin, A., Naveiras, M., & Salas, J. (2024). Differential and functional response time item analysis: An application to understanding paper versus digital reading processes. *Journal of Educational Measurement*, *61*, 219–251.

Coleman, J. S. (1968). Equality of educational opportunity. *Integrated Education*, *6*(5), 19–28.

Darling-Hammond, L. (2015). *The flat world and education: How America's commitment to equity will determine our future*. Teachers College Press.

Domingue, B., Kanopka, K., Stenhaug, B., Soland, J., Kuhfeld, M., Wise, S., & Piech, C. (2021). Variation in respondent speed and its implications: Evidence from an adaptive testing scenario. *Journal of Educational Measurement*, *58*, 335–363.

Domingue, B., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradović, J., Piech, C., Porter, T., Consortium, P. iLEAD, Soland, J., Weeks, J., Wise, S. L., & Yeatman, J. (2022). Speed–accuracy trade-off? not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *Journal of Educational and Behavioral Statistics*, *47*(5), 576–602.

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

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill:

Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608.

Guo, H., & Ercikan, K. (2020). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, *26*(5-6), 302–327.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Heitz, R. P., & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General*, *136*(2), 217.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139.

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika*, *87*, 725–748.

Kanopka, K., & Domingue, B. (2022). A position sensitive IRT mixture model. https://doi.org/10.31234/osf.io/hn2p5

Leary, L. F. & Dorans, N. J. (1982). The effects of item rearrangement on test performance: A review of the literature. *ETS Research Report Series*, *1982*(2), i–23.

Linn, M. C., De Benedictis, T., DeLucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "I don't know" really mean? *Journal of Research in Science Teaching*, *24*(3), 267–278.

Lounkaew, K. (2013). Explaining urban-rural differences in educational achievement in Thailand: Evidence from PISA literacy data. *Economics of Education Review*, *37*, 213–225.

Mutak, A., Krause, R., Ulitzsch, E., Much, S., Ranger, J., & Pohl, S. (2024). Modeling the intraindividual relation of ability and speed within a test. *Journal of Educational Measurement*, *61*, 378–407.

Niederle, M., & Vesterlund, L. (June 2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, *24*(2), 129–44. https://doi.org/10.1257/jep.24.2.129.

OECD. (2019). *PISA 2018 Science Framework*. Paris: OECD Publishing. https://doi.org/10.1787/f30da688-en.

OECD. (2019a). PISA 2018 Reading Framework. In *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing. https://doi.org/10.1787/5c07e4f1-en.

OECD. (2019b). *Released items from the PISA 2018 computer-based reading assessment*. Paris: OECD Publishing.

Ranger, J., Kuhn, J.-T., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 15–38.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2014). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. A. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 507–525). Routledge.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*(5), 284–294.

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2019). Gender achievement gaps in US school districts. *American Educational Research Journal*, *56*(6), 2474–2508.

Schleicher, A. (2019). PISA 2018: Insights and interpretations. Paris: OECD Publishing.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A., EORTC Quality of Life Group, &

Quality of Life Cross-Cultural Meta-Analysis Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295.

Shear, B. R. (2023). Gender Bias in Test Item Formats: Evidence from PISA 2009, 2012, and 2015 Math and Reading Tests. *Journal of Educational Measurement*, *60*, 676–696.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, *66*, 487–518.

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

Soland, J. (2018b). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12), 1–26.

Solheim, O. J., & Lundetræ, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults?—A study based on Nordic results in PIRLS, PISA and PIAAC. *Assessment in Education: Principles, Policy & Practice*, *25*(1), 107–126.

Solheim, O. J., & Lundetræ, K. (2020). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults?—A study based on nordic results in PIRLS, PISA and PIAAC. *Assessment of Reading in International Studies*, *25*, 107–126.

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.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Wang, S., & Chen, Y. (2020). Using response times and response accuracy to measure fluency within cognitive diagnosis models. *Psychometrika*, *85*(3), 600–629.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Yousefpoori-Naeim, M., Bulut, O., & Tan, B. (2023). Predicting reading comprehension performance based on student characteristics and item properties. *Studies in Educational Evaluation*, *79*.

Appendix 1: Score Differences Adjusted for Capacity Differences: Simulation and Empirical Results

Simulation Results

This section presents results for accuracy differences due to speed ($\widehat{A_s}$), that is, score differences adjusted for capacity differences. We demonstrate the results here to complete the intuition of the simulation. Difference in score between groups (A) are simulated to be a function of latent speed (μ_τ), latent capacity (μ_θ), and slope of the CAF (b_t). The top panel in Figure A1 considers change in group difference due to speed, $\widehat{A_s}$ over 100 values of capacity μ_θ. As expected, group differences due to speed ($\widehat{A_s}$) is higher for the slower group (red) when CAF is upward sloping ($b_t > 0$ in the *top left* panel); when CAF is downward sloping ($b_t < 0$ in *top right* panel), the adjusted difference $\widehat{A_s}$ is lower for the slower group. Note that $\widehat{A_s}$ is not affected by change in capacity μ_θ, it only changes with μ_τ.

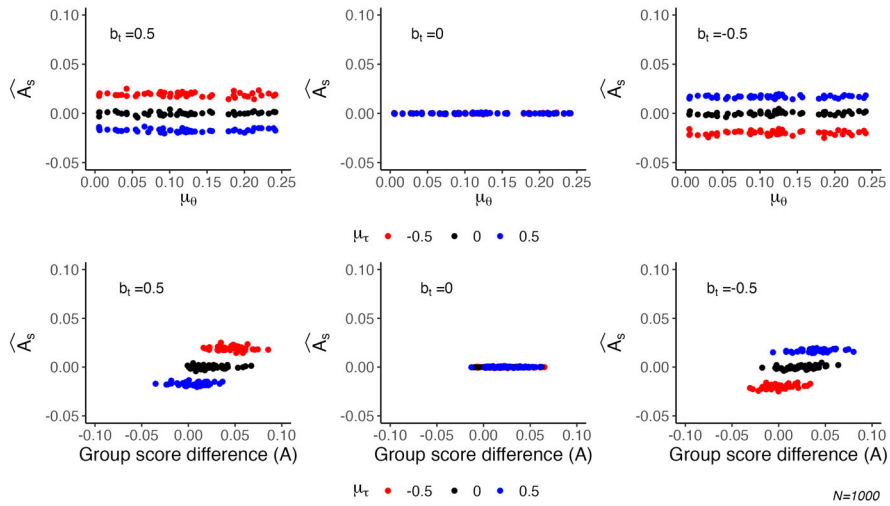

Figure A1. Results from simulation study for 100 values of μ_θ for each set of simulation conditions (N, μ_τ), when $N = 1,000$. *Top* panel shows the association between μ_θ and the adjusted score $\widehat{A}_s$, for $b_t \in \{.5, 0, -.5\}$. Note that after adjusting score for differences in capacity, the score difference $\widehat{A}_s$ is associated with μ_τ, but not with μ_θ; that is, any difference in scores is due to difference in latent speed. In the *top left* panel, CAFs are upward sloping ($b_t > 0$), and hence, groups with slower speed (red) have the higher score. In the right panel, red group has slower speed and lower score as CAFs are downward sloping ($b_t < 0$). The *bottom row* column shows the association between Observed difference A and $\widehat{A}_s$. Note that with $\widehat{A}_s$, adjusted scores are the same for all values of A as capacity differences have been adjusted, similar to the top panel.

We see this relationship in the bottom panel which compares the observed group score A with scores adjusted for capacity differences ($\widehat{A}_s$). The results show that variation in A estimates translates to variation due solely to μ_τ when we consider $\widehat{A}_s$. This is as expected given that $\widehat{A}_s$ indexes speed-based differences; differences between similar colored points are minimized as they are due solely to μ_θ (to which $\widehat{A}_s$ is insensitive). Similar to the top row, $\widehat{A}_s$ is centered around a constant value given the underlying μ_τ at all values of group score difference A. In the *bottom left* panel, when CAF is upward sloping, the slower red group ($\mu_\tau < 0$) are centered around a higher constant $\widehat{A}_s$, and the faster blue group around a lower $\widehat{A}_s$. The direction is reversed when CAFs are downward sloping in the right panel. Further, if CAFs were flat, there would be no systematic change from the adjustment ($b_t = 0$ in the middle panel).

Empirical Results

This section discusses adjusted group score differences due to speed ($\widehat{A}_s$) as compared to observed differences for PISA scientific literacy and fluency. Figure A2 shows that when males are the higher performing group in science, group differences ($\widehat{A}_s$) are below 0. Note that males are the higher performing, faster speed group that

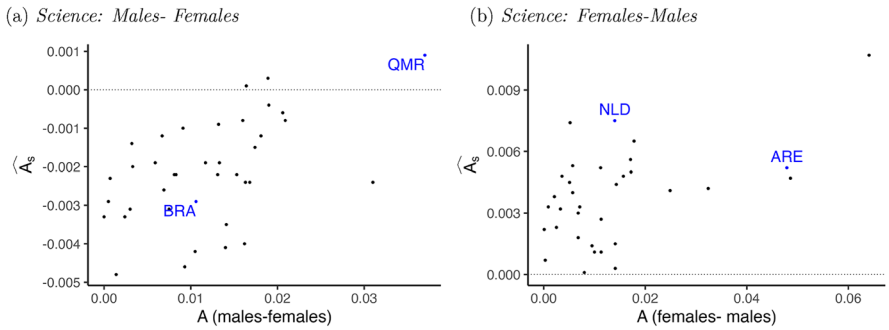

Figure A2. **PISA Scientific literacy**: Results across countries for differences due to speed, that is, differences adjusted for capacity. *Left*: Adjusted differences when males score more than females. *Right*: Adjusted differences when females score more than males

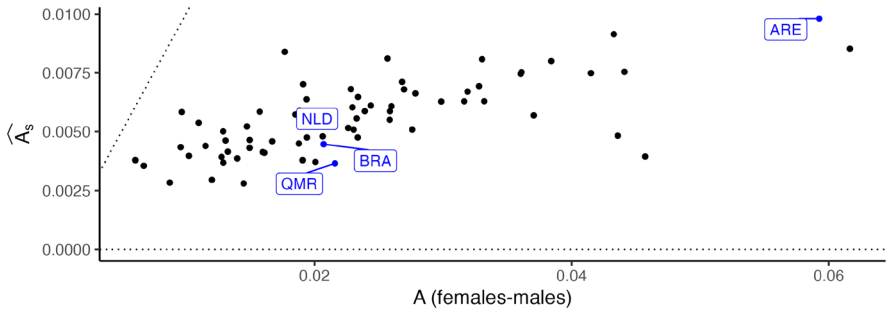

Figure A3. PISA Fluency: Results across countries for differences due to speed, that is, differences adjusted for capacity. Females scored more than males across countries in fluency

spend less time on items. In the simulation (Figure A1), this is the bottom left panel, where males correspond to the blue dots. Females correspond to the red dots in the panel, where adjusted score increases because the higher performing group has less speed and spends more time on items.

For reading fluency (Figure A3, females are the higher performing group that have more speed (less time on items) when CAFs are downward sloping. This corresponds to the bottom right panel in Figure A1, where females now correspond with the blue dots. Note that adjusted differences increase for most countries.

Appendix 2: Simulation Results: Sensitivity to Choice of Parameters

Number of Respondents: $N = 10{,}000$

Group score differences are adjusted for speed ($\widehat{A_s}$) and capacity ($\widehat{A_c}$) for $N = 10,000$. These results are similar to the results for $N = 1,000$, but offer more precision. These results show detail related to, for example, the degree to which increased sample results in more precise estimates of $\widehat{A_c}$ quantities but suggests a similar set of

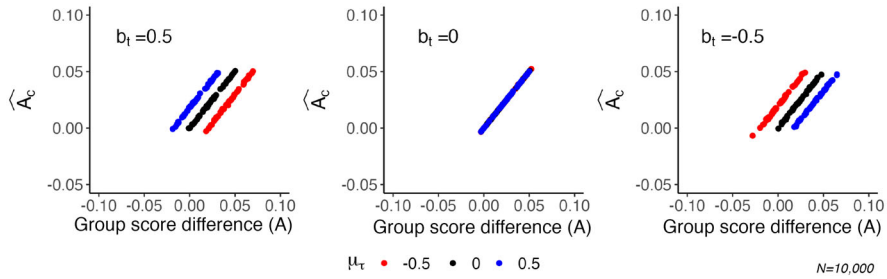

Figure B1. **Results from simulation study for 100 values of μ_θ for each set of simulation conditions** (N, μ_τ), **when** $N = 10,000$. Graphs show relationship between observed group differences (A) and true group capacity difference μ_θ. Similar to $N = 1,000$, observed group differences are separated by μ_τ at the same capacity. The variance is estimates is lower for larger N

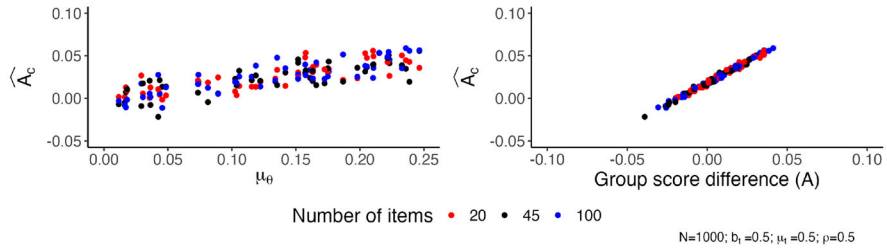

Figure B2. **Results from simulation study for 100 values of μ_θ, when number of items $\in \{20, 45, 100\}$, for each set of simulation conditions** (N, μ_τ). *Left*: Graph shows the association between μ_θ and the adjusted score $\widehat{A_c}$, with the colors showing variation in number of items. The *right* column shows the association between Observed difference A and $\widehat{A_c}$. There is no variation in results based on number of items

facts about their capacity to distinguish between variation in observed accuracy due to μ_τ versus μ_θ. Based on these results, we argue that the $\widehat{A_c}$ estimates are effectively able to disentangle the role of speed and capacity underlying naive A estimates for a wide range of sample size (Figure B1).

Sensitivity to Number of Items and Capacity-Speed Covariance

Figure B2 shows that the simulation results are not sensitive to choice of number of items between 20, 45, and 100 items. This is shown by the overlapping red, blue, and black dots when $\widehat{A_c}$ is mapped against ability (μ_θ) and observed group score difference A. Similarly, Figure B3 shows that results are not sensitive to the choice of covariance between latent capacity and latent speed (ρ). The figure holds all other parameters including covariance for group 2 constant $(\rho_2 = 0)$, while varying covariance for group 1 $(\rho_1 \in \{0, .25, .5\})$.

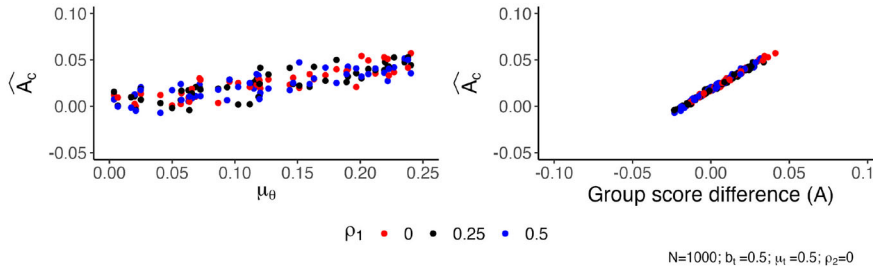

Figure B3. **Results from simulation study for 100 values of μ_θ, when latent capacity-latent speed covariance for group 1 $\rho_1 \in \{0, .25, .5\}$, for each set of simulation conditions** (N, μ_τ). *Left*: Graph shows the association between μ_θ and the adjusted score $\widehat{A_c}$, with the colors showing variation in ρ. The *right* column shows the association between Observed difference A and $\widehat{A_c}$. There is no variation in results based on ρ

Appendix 3: Country-Level Descriptive Results

Please find country- and subject-level descriptive results in the attached Supplementary Index file, including number of respondents, percent of female respondents, observed score differences (A), and adjusted differences ($\widehat{A_c}$).

Appendix 4: Empirical Results: Brazil, Moscow, Netherlands, United Arab Emirates

PISA Scientific Literacy

This section shows country specific CAFs and response time distributions for highlighted countries. These countries were selected to offer a glimpse into how different permutations of the method were working. In the left panel of Figure D1 below, females are the lower capacity group in Brazil (lower CAF) which spend more time on items. Adjusting for time use hence increases group differences. In Moscow (right panel), females were the lower performing group that spent less time on items. In other words, males were the higher capacity group that spent more time on items. Adjusting for time use differences reduces group differences.

In Figure D2, females are the higher capacity group for both Netherlands and United Arab Emirates (UAE) for scientific literacy. Females are also the group that spends more time on items. Adjusting for time use differences reduces group differences.

PISA Reading Fluency

In reading fluency, we note that group CAFs intersect for Netherlands, UAE, and Brazil (Figure D3). CAFs for females rises above the CAFs for males for response time between –2 and 1 log seconds for Netherlands, between –4 and 1 log seconds for UAE, and between –3 and 1 log seconds for Brazil. Note that CAFs for females are more curvilinear; that is, returns to time both increases and decreases more sharply

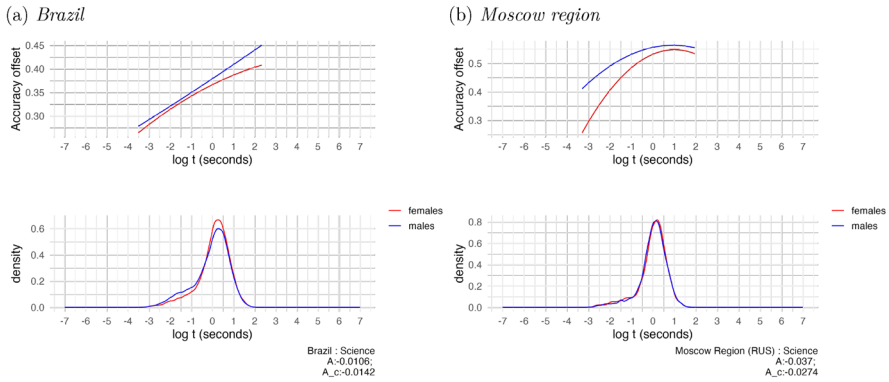

Figure D1. **PISA Scientific literacy**: Results from two countries where males scored more than females. Score difference between groups decreases for Moscow but increases for Brazil.

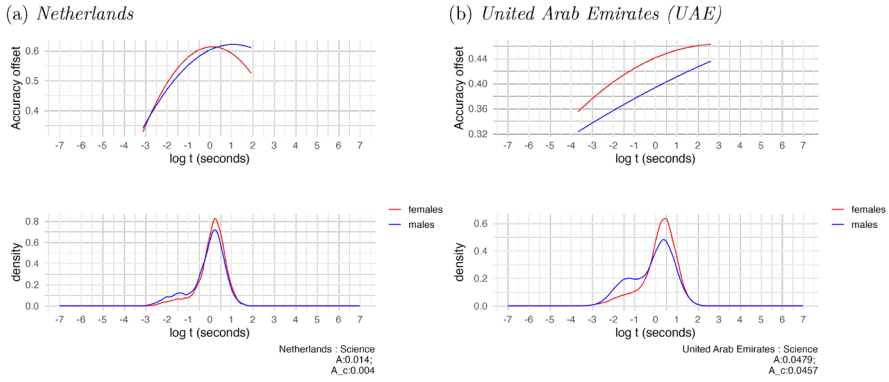

Figure D2. **PISA Scientific literacy**: Results from two countries where females scored more than males. Score difference between groups decreases for both Netherlands and UAE.

for females. CAFs turn downward sloping between –1 and 0 log seconds for all countries. Note that this is the part of CAFs where the female CAF lies above the male CAF. Adjusting for time would shift the time distribution for females to the right; that is, they would be spending more time on items in the region of downward sloping part of CAFs. This implies that adjusting for time distribution would reduce group differences, as females are the higher performing group with downward CAFs and spending less time on items. This is what we observe, as group differences reduce from 2.12% to 1.73%.

For Moscow Region of Russia, CAF for females is always above the graph for males, and females spend less time on times. Adjusting for speed reduces group differences from 2.16% to 1.82%, suggesting that the impact of the adjustment was in the downward part of the CAF.

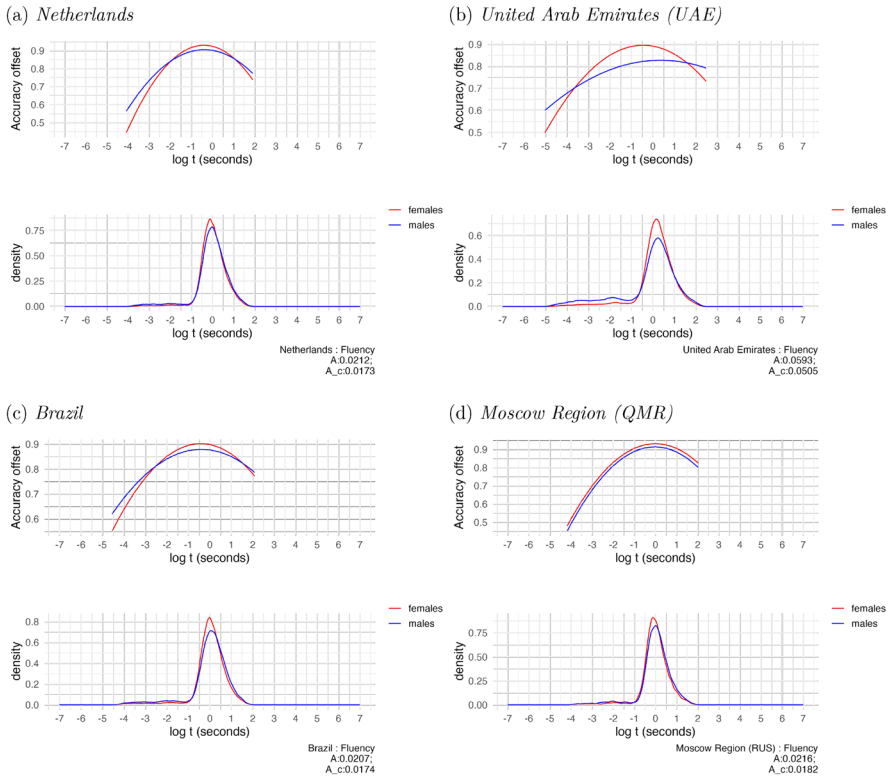

Figure D3. **PISA reading fluency**: Results from two countries where females scored more than males. Score difference between groups decreases for both Netherlands and Moscow Region (QMR).

Authors

Radhika Kapoor is a PhD candidate, Stanford Graduate School of Education, 482 Galvez Mall, Stanford, CA 94305–3096, rkap786@stanford.edu. Her primary research interests include psychometric methods, international assessments, and machine learning.

Erin Fahle is the co-director of the Educational Opportunity Project, Stanford Graduate School of Education, 482 Galvez Mall, Stanford, CA 94305–3096, efahle@stanford.edu. Her primary research interests include education policy.

Klint Kanopka is an Assistant Professor of Applied Statistics at New York University, Steinhardt School of Culture, Education, and Human Development, in the Department of Applied Statistics, Social Science, and the Humanities, 246 Greene Street, Floor 3 New York, NY 10003; klint.kanopka@nyu.edu. His primary research interests concern the intersection of psychometric methods and machine learning.

David Klinowski is a Visiting Assistant Professor of Marketing and Business Economics at the Katz Graduate School of Business, University of Pittsburgh, 3950 Roberto & Vera Clemente Drive, Pittsburgh, PA 15260. His primary research interests include experimental economics and the economics of gender.

Ana Trindade Ribero is a Senior Researcher at the National Student Support Accelerator at Stanford University, 520 Galvez Mall, Stanford, CA 94305, anactr@stanford.edu. Her pri-

mary research interests include behavioral economics, natural language processing, and education policy.

Ben Domingue is an associate professor, Stanford Graduate School of Education, 482 Galvez Mall, Stanford, CA 94305–3096, bdomingue@stanford.edu. His primary research interests include psychometrics.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information