I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17734

Time Well Spent? The Role of Test Effort in Explaining Achievement Gaps

Lex Borghans
Ron Diris
Mariana Tavares

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17734

Time Well Spent? The Role of Test Effort in Explaining Achievement Gaps

Lex Borghans
Maastricht University and IZA

Ron Diris
Leiden University and IZA

Mariana Tavares
Forward College

FEBRUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Time Well Spent? The Role of Test Effort in Explaining Achievement Gaps*

In this paper, we identify the contribution of differences in test effort to gender gaps and socioeconomic gaps in achievement. We leverage question response time and random question order to obtain causal estimates of the effect of student effort on performance. Subsequently, we evaluate how differences in performance change when students would have made equal time investments. We find that effort explains around 25 percent of the socioeconomic gap in math and reading. For gender, correcting for effort closes around 18 percent of the reading gap while it increases the advantage of boys in math. Looking at average achievement, gender differences in effort can explain 49 percent of the gender achievement gap. We also show that the returns to response time are strongly underestimated by fixed effects models.

JEL Classification: I21, I24

Keywords: education economics, achievement gaps, student effort, instrumental variables

Corresponding author:
Ron Diris
Department of Economics
Leiden University
Steenschuur 25
2311 ES Leiden
The Netherlands
E-mail: r.e.m.diris@law.leidenuniv.nl

* The authors thank Sergio Correia for his insights on Statalist regarding statistical tests across specifications. The authors thank conference participants at the European Association of Labour Economics 2021 and the European Society for Population Economics 2023 in Belgrade for their comments, as well as seminar attendees at the NIPE seminar series 2021/22 organised by the *Economic Policies Research Unit* from the *School of Economics and Management of the University of Minho*.

1 Introduction

The fact that educational achievement systematically differs between certain groups of students is a central topic in educational research and practice. Achievement gaps by gender and by socioeconomic status (SES hereafter) have received special attention, because they are argued to be a key factor in reproducing inequalities in socioeconomic outcomes and representation in adult life (Bedard & Cho, 2010; Hanushek *et al.*, 2019). To effectively remediate such gaps, understanding what they reflect is crucial. Typically, analysis and discussion of achievement gaps perceive these as reflecting differences in cognitive ability or acquired knowledge, and remediation of achievement gaps is consequently also targeted at these aspects. However, recent evidence shows that test scores are also substantially driven by student effort during the test (Zamarro *et al.*, 2016; Gneezy *et al.*, 2019), and that such effort typically differs across groups (Akyol *et al.*, 2021; Borghans *et al.*, 2024).[1] These observations beg the question to what extent gender and SES achievement gaps can be attributed to differences in effort between these groups.

In this paper, we causally identify to what extent gender and SES achievement gaps are driven by differences in test-taking effort. We use data from 14,981 9th grade students in the Netherlands who take a low-stakes computerized test measuring IQ, math and reading achievement. The test data have two key features: 1) they contain detailed timestamps, allowing us to infer question response time as a measure of effort, 2) question order is randomly assigned across students, allowing us to leverage exogenous variation in response time to estimate the causal effect of effort on achievement.

Response time is an endogenous variable, as students may take more or less time depending on the difficulty of the question and/or their own ability level. Crucially, this endogeneity bias is not solved by including question and individual fixed effects, because this does not take into account that they may interact: particular questions may be easier or more difficult for a particular student, which can influence the effort s/he will put forth on that question. To tackle this source of

[1] For a more general discussion on the importance of effort for school achievement, see, e.g., Borghans *et al.* (2008a).

endogeneity, and to causally identify the effect of student effort on performance in IQ, math and reading questions, we take advantage of the fact that students invest less time per question as these move towards the end of the test (Tavares, 2022). Due to the randomness of the question position in the test, we can isolate the exogenous variation in response time for a given question that stems from test progression.

We find that extra response time has no effect on performance in the IQ test, but has a sizable effect on achievement in the math and reading tests. The IV estimates are substantially larger than the naive OLS estimates and estimates including student fixed effects. This shows that a negative bias arises in the latter from students taking more time when a question is particularly difficult to them. We find that ten seconds extra response time causally increases the probability of a full score in math and reading by approximately 3 percentage points. Applying these causal estimates, we find that the gender achievement gap in reading closes by 18.4 percent when there would be no differences in response time. The SES achievement gaps in reading and math similarly close by 27 percent and 21 percent, respectively. Conversely, the gender gap in math increases, since girls perform worse on math while putting in more effort. When looking at average achievement across math and reading, we find that student effort explains 49 [24] percent of the achievement gap across gender [SES].

This paper brings together literature about the gender and SES achievement gaps and literature about the role of effort in test taking. Studies on gender and SES achievement gaps are abound in empirical research, as these are highly relevant to inequalities in society. Many studies aim to discern the underlying reasons for the identified gaps. SES gaps have been highlighted as far back as the Coleman report (Coleman, 1968) and have since particularly focused on the relative contributions of (early) family circumstances (Heckman, 2000) and school quality (Jennings *et al.*, 2015; Reardon, 2016). The role of effort differences across SES groups has received little attention, however.

For gender gaps, it is typically identified that female students perform substantially better in reading, while math gender gaps tend to be more inconsistent across countries but, on average, in

favour of male students (Ceci *et al.*, 2014; Breda & Napp, 2019; Ellison & Swanson, 2023). Analyses of the explanatory factors for these gaps in achievement have focused on many factors, including the role of biological differences (Wilder & Powell, 1989), comparative advantage (Breda & Napp, 2019), bias in class-level streaming (Bedard & Cho, 2010), stereotype threat (Spencer *et al.*, 1999), parental expectations (Bhanot & Jovanovic, 2005), and noncognitive skills (Cornwell *et al.*, 2013).[2] Our study relates mostly to the latter. Nonetheless, we consider effort as a separate concept that may be influenced by non-cognitive skills, and is a measure of direct behaviour rather than of an underlying trait or skill. Additionally, research shows that student effort during a test predicts test scores and that country differences in effort affect country rankings on standardized tests such as PISA (Zamarro *et al.*, 2016; Gneezy *et al.*, 2019; Akyol *et al.*, 2021). Moreover, studies have identified that such effort predicts future school outcomes (Xiong *et al.*, 2011; Hernández & Hershaff, 2015).

A central question in this literature is how to measure test effort. Setzer *et al.* (2013) reflect on the capacity of response time to capture effort and argue in favour of this measure compared to traditional effort measures.[3] Alternative measures are self-reported effort, non-response rates, careless answering behaviour, and performance decline. The first is subjective and subject to reference bias, non-response rates are ill-suited for multiple-choice questions, while careless answering is better suited for surveys. Hence, response time provides an objective measure, comparable across students, that can be used across different types of tests. Previous literature on response time typically uses it to binarily classify responses as non-effortful *vs.* solution behaviour, and assess overall test motivation or to analyse and improve item validity. This study is, to the best of our knowledge, the first that explores the (causal) role of response time in explaining achievement gaps.

This study also relates to research on performance decline across tests (Borghans & Schils, 2018; Zamarro *et al.*, 2016; Borgonovi & Biecek, 2016; Brunello *et al.*, 2018; Balart & Ooster-

[2] Fryer Jr & Levitt (2010) explore several of these potential explanations (differential investments, parental expectations, biased tests) and find little evidence for each of them.

[3] The OECD includes response time among their effort and motivation indicators in PISA questions (OECD, 2019).

veen, 2019). This literature shows that female and high SES students are better able to sustain initial levels of performance compared to their male and low SES counterparts. This literature also analyses effort in relation to achievement gaps (specifically by gender), but focuses on other dimensions of effort. In particular, while their focus is on an output measure (performance decline) ours is on an input measure (response time). Second, our focus is not on effort decline but on its level. Our study thus comes with a different aim and different potential implications. Studies such as Balart & Oosterveen (2019) hold important implications for the sensitivity of measured gender gaps in achievement across tests of different length. Our results predominantly pertain to how effort (causally) contributes to achievement gaps in a given test, and what share of those gaps it can explain.

In summary, we contribute to the literature by quantifying the causal effect of effort exerted during a test, and using those estimates to explain the contribution of effort to achievement gaps in math and reading. Importantly, previous work has assumed that the mismatch between student knowledge and question content has negligible implications for delivered effort levels. We employ an identification strategy that addresses this concern. We show that IV estimates are statistically different from estimates from an OLS model with fixed effects, which rely on this assumption and are thus biased. By doing so, we provide evidence on a potential determinant of achievement gaps that has received little attention in this vast literature, and that may thus impact our understanding of these gaps, and how they can potentially be remediated.

The remainder of the paper is organized as follows. Section 2 describes the data set and presents descriptive statistics. Section 3 focuses on methodology and identification strategy. Section 4 reports and discusses our results, while Section 5 concludes.

2 Data and Descriptive Statistics

We use data from the Onderwijsmonitor Limburg (OML hereafter), an ongoing regional education monitor that gathers data since 2009 in Limburg, a province of the Netherlands.[4] This is a cooperative project between Maastricht University and both primary and secondary schools, with a participation rate above 90 percent (Hirsch, 2017). We use testing data from students in 9th grade (lower secondary school). Dutch primary education ends in 6th grade, after which students are assigned to secondary school tracks that lead to specific labour market qualifications. There are four main secondary school tracks: two are vocationally oriented (lower and upper vocational) and two are academically oriented (general education and pre-university education).

OML administers a computerized test to 9th grade students, every two years, with the purpose of research and providing feedback to schools. The test is administered in the classroom, within a 50 minutes window and under teacher supervision. It comprises IQ, math and reading questions and has three versions, depending on secondary school track: lower vocational students take *test version 1*, upper vocational students take *version 2* and academic-oriented tracks take *version 3*. Table 1 shows how students are distributed across test versions.

In addition to the different versions that make sure that questions fit with the respective track curriculum, students are randomly assigned to four different "routes". Everyone starts with the IQ block, after which students either get three blocks of math questions (25 percent of students), three blocks of reading questions (25 percent), two blocks of math and one block of reading questions (25 percent), or two blocks of reading and one block of math questions (25 percent). The four possible routes are thus: *IQ-MA-MB-MC*, *IQ-RA-RB-RC*, *IQ-MA-RA-RC* and *IQ-RA-MA-MC*. Table 1 shows by subject how many students have made the particular test.

The test software ensures transition between the blocks either when all questions are completed or a time limit is reached. When a student is still in the first block after 10 minutes, s/he is transitioned to the next block as soon as they submit an answer to the current question. For the

[4] OML is part of a larger cooperative program, Educatieve Agenda Limburg, between Maastricht University and school boards in primary and secondary and higher professional education in Limburg. Limburg is one of the twelve provinces of the Netherlands, covering the southeastern part of the country.

Table 1: Descriptive Statistics: average, standard deviation and frequency

	IQ	Math	Reading	All
Total number of students	7,400	10,980	11,375	14,980
in 2012	3,136	4,657	4,939	6,399
in 2014	1,544	2,299	2,336	3,081
in 2016	2,720	4,024	4,100	5,470
Test version 1	1,488	2,231	2,292	3,010
Test version 2	2,161	3,144	3,283	4,304
Test version 3	3,751	5,605	5,800	7,636
Number of student-question pairs	137,192	172,172	75,415	384,779
% Female students	0.51 (0.50)	0.52 (0.50)	0.52 (0.50)	0.52 (0.50)
% High SES students	0.39 (0.49)	0.39 (0.49)	0.42 (0.49)	0.40 (0.49)
Score	57.03 (49.50)	44.73 (48.19)	66.13 (32.63)	53.31 (46.81)
Response time	23.24 (17.80)	52.98 (51.36)	77.92 (65.80)	47.26 (50.46)

Note: The first panel displays frequencies (e.g.: number of students in the sample), whereas the second shows sample averages and standard deviations in parenthesis. Note that the a student may have been tested in more than one subject in the same year.

second and third blocks, these limits are set at 15 minutes, whereas the fourth and last block has no fixed time limit. Blocks are usually finished, especially in reading (98 percent of all cases) and math (75 percent of all cases, with another 10 percent finishing all but one question). For the IQ block, only 41 percent of students answer all questions.

Additionally, questions are fully randomized inside each block.[5] There are thus two forms of randomization that create variation in question order. In the main empirical approach, we only use the within-block variation, as we consider it more reliable. Variation in test routes does not only determine whether math questions are received early or late in the test but also whether they are preceded by reading questions or not, which may have an independent effect. Appendix Table E1 attests the random nature of question order within blocks in our sample.

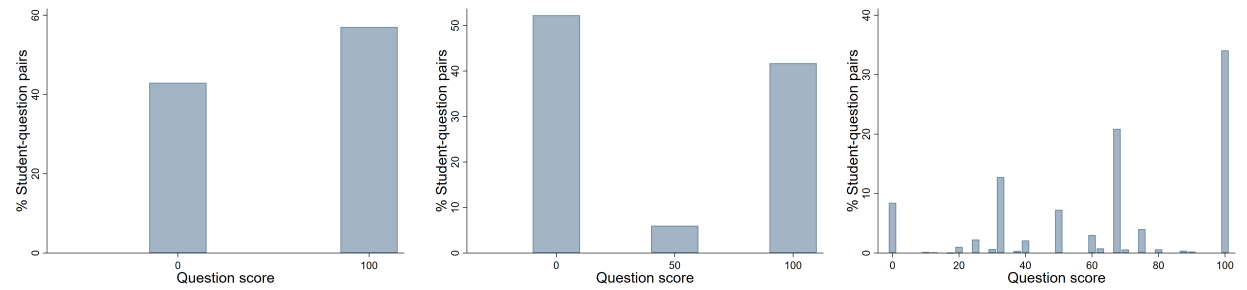
The data collected during the test include question scores and timestamps for each question-student pair. Timestamps are collected by "screen": the software records whenever a new screen with a new question appears, and when the answer to that question is submitted. In some instances, there are multiple questions on the same screen (typically for reading when there are several questions about the same text). As we cannot deduce timestamps for each, we aggregate scores by screen. We consider everything that appears jointly on one screen as a single *question*, while we refer to subquestions within one screen as *items*. Question scores thus correspond to the percentage of correctly answered items within each question for each student. All IQ questions are single-item as are 90 percent of math questions. Reading questions are all multi-item. On average, each student answers 18 IQ questions, 16 math questions and 8 reading questions. Figure 1 shows the distribution of question scores for each subject in our sample. The average score equals 57 percent for IQ questions, 45 percent for math questions and 66 percent for reading questions.

Timestamps are available for tests administered in 2012, 2014 and 2016.[6] Response time of student i to question j, in seconds (r_{ij}), corresponds to the difference between the two timestamps recorded for each question-student pair. The first is the moment when question j (with all its items) is displayed on the computer screen of student i, whereas the second corresponds to the moment when student i submits an answer to question j.

[5] For half of the students, IQ questions are grouped within sub-blocks (according to type). Given our identification strategy, we drop these student-question pairs, which explains the lower number of students under IQ in table 1. Test versions represent groups of questions and some questions overlap across versions. For a detailed description of the test, please see Tavares (2022).

[6] As shown in Table 1, the number of students with testing data is substantially lower in 2014. Because of planning difficulties, only around half of all schools were able to administer the test. Results are very similar when we exclude the 2014 cohort.

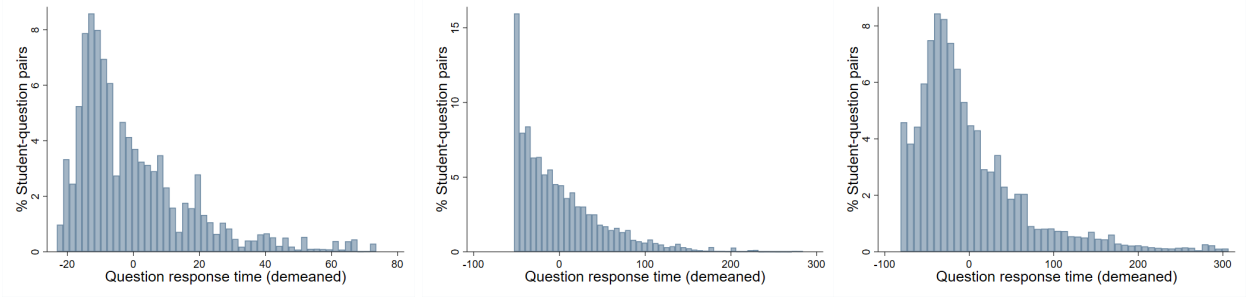
Figure 1: Distribution of question score



(a) IQ (b) Math (c) Reading

We use the median absolute deviation method, with a threshold of 3, to identify right-tail outliers in the distribution of response time. Response times of observations classified as outliers are winsorized, and thus re-coded to the corresponding outlier value.[7] After this adjustment, the average response time equals 23.2 seconds for IQ questions, 53.0 seconds for math questions and 77.9 seconds for reading questions. Figure 2 shows the distribution of question response time for each subject in our sample, after winsorizing outliers and demeaning question response time within each subject.[8]

Figure 2: Distribution of question response time



(a) IQ (b) Math (c) Reading

The second part of the analysis is aimed at explaining gender and SES achievement gaps.

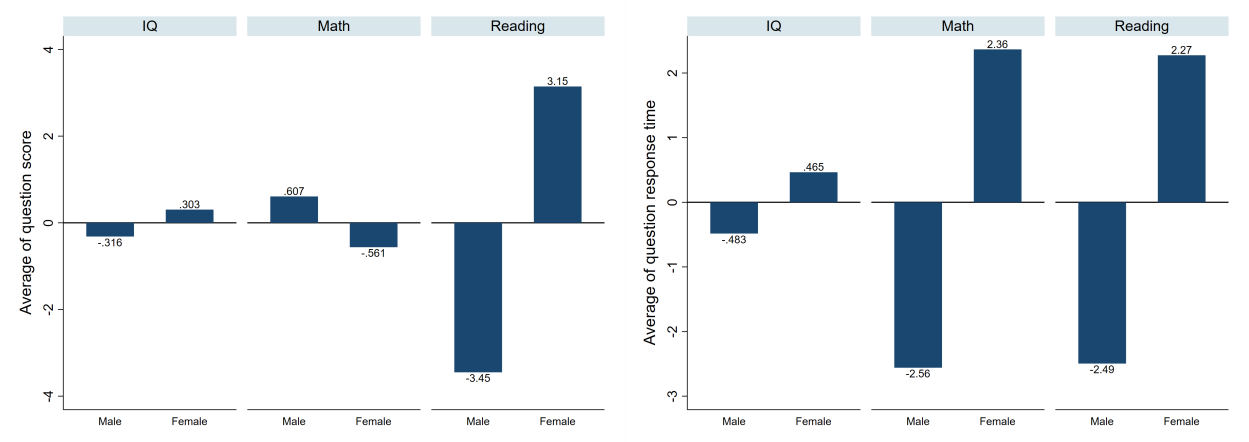
[7]Particularly, an observation (student-question pair) is classified as an outlier if its response time exceeds the average response time plus 3 times the median absolute deviation from the question's median response time, $r_{ij} > \bar{r_j} + 3 \times median(|r_{ij} - median(r_{i,\bar{j}})|)$ for fixed j, where $\bar{r_j} = \sum_{i=1}^{n_{\bar{j}}} \frac{r_{ij}}{n_{\bar{j}}}$. Response time of observations deemed as outliers is recorded to $\bar{r_j} + 3 \times median(|r_{ij} - median(r_{i,\bar{j}})|)$. Dropping outliers does not change our results. It increases point estimates of the effect of response time on question score.

[8]Figure 2 shows that there is a relatively large set of short answers for math items. Excluding these does not qualitatively change our results.

Student gender is taken from the school administration and SES is collected from student and parent questionnaires. Students are classified as coming from a high socioeconomic background when at least one of the parents has finished tertiary education (either university of applied sciences or university).

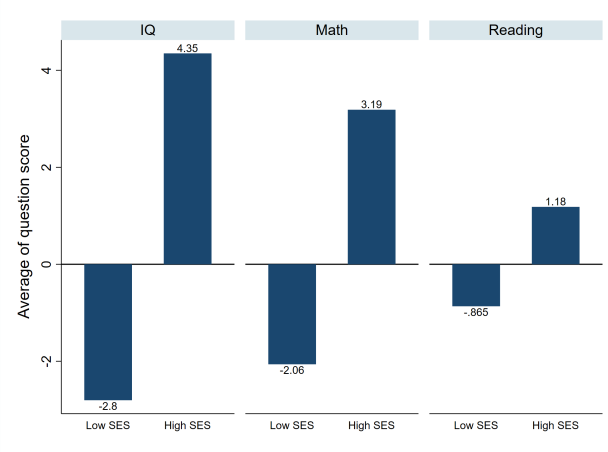
Figure 3 shows descriptive information on gender and SES gaps in performance and in question response time. Panel (a) suggests that female students have higher average scores in reading than male students, but lower scores in math. Average IQ scores are very similar by gender. Response time averages are higher for female students than male students across all subjects. They are of similar size for math and reading, and lower for IQ questions. Panel (b) shows that high SES students have higher average scores as well as higher response times. Score differences are more prominent in IQ than in math and reading. Response time differences are largest for math, both absolutely and relatively, followed by reading and IQ. Figure 3 thus shows that there is variation in response time by gender and SES, and therefore that such differences could potentially explain part of achievement gaps. To quantify this, we first need to establish whether response time causally contributes to achievement and, if so, by how much.

Figure 3: Average differences in performance and effort across groups

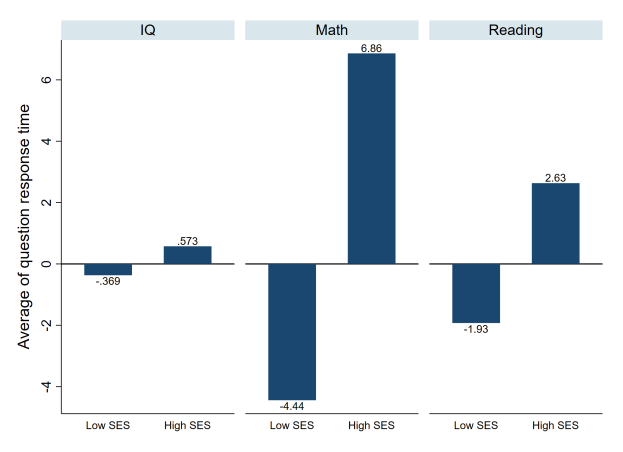


(a–1) Gender: score difference

(a–2) Gender: time difference



(b–1) SES: score difference



(b–2) SES: time difference

Note: Both test score and response time are demeaned within subject. Question score in the figure excludes the effect of school year and question characteristics. It corresponds to the (demeaned) residuals of regressing each question-student score on school year dummies and question fixed effects.

3 Methodology

First, we aim to identify the causal effect of response time on question scores. Second, we aim to quantify the implications of response time differences for gender and SES achievement gaps in IQ, math and reading tests. The methodology adopted regarding the first is covered in subsection 3.1, whereas the second is presented in subsection 3.2.

3.1 Estimating the causal effect of response time

We first specify the typical approach in estimating the effect of response time on scores. Equation 1 denotes question score for student i and question j ($score_{ij}$) as a function of demeaned response time ($\tilde{r}_{ij}$) (in seconds), question fixed effects (α_j) and student fixed effects ($\ddot{\alpha}_i$).[9] The error term e_{ij} follows the standard assumptions. Standard errors are clustered at the student-level, since individual scores are exposed to unobserved factors that are individual-specific.[10]

[9] We use a linear specification as our main focus is not on the behaviour of observations in the tails of the distribution and we believe this is more suitable given our instrumental variable. Moreover, after including test controls and fixed effects, the relation between exogenous response time and performance approaches a straight line. Results are similar if higher order polynomials of response time are included, namely squared and cubic terms. We elaborate on this issue in Section 4.3.

[10] Clustering at both question and student levels does not affect the significance of our results.

$$score_{ij} = \alpha_j + \ddot{\alpha}_i + \beta_0 \tilde{r}_{ij} + e_{ij} \tag{1}$$

The coefficient of interest in equation 1 is β_0. It reflects how the average probability to answer question j correctly changes when student i devotes more time to question j, all else equal. We also estimate equation 1 without both student and question fixed effects and with only question fixed effects. This allows us to see how these typical correction approaches matter for the identified effect of response time on score.

Question fixed effects absorb question features that are student and order-invariant (such as difficulty), whereas student fixed effects control for student characteristics that are question and order-invariant (such as general ability and motivation to engage with the test). There may be, however, student-question factors that explain both question score and response time, in particular the mismatch between question content and student knowledge. As students may adjust their response time if they perceive the question as harder or easier, also net of average question difficulty and average student performance, we suspect that response time may suffer from endogeneity as a predictor of student performance.

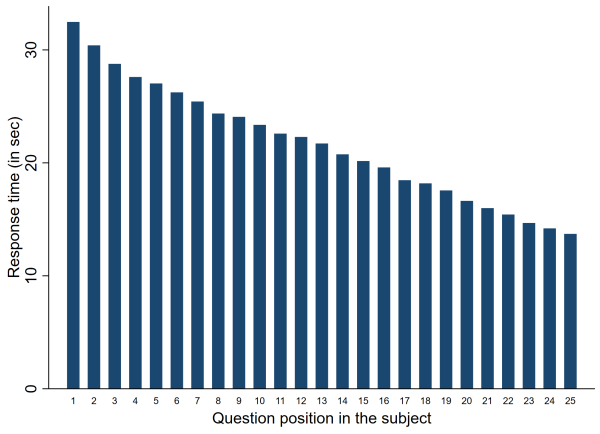
In order to tackle this identification problem, we take advantage of random question order. It ensures that question (and student) characteristics are independent of question position. In this way, we capture the exogenous variation in response time that stems from test progression. We estimate an IV model where the endogenous variable, $\tilde{r}_{ij}$, is instrumented with the (demeaned) relative question position in the test block ($\widetilde{Q}^p_{ij}$). The first stage equation thus becomes:

$$\tilde{r}_{ij} = \alpha_j + \ddot{\alpha}_i + \gamma_0 \widetilde{Q}^p_{ij} + \dot{e}_{ij} \tag{2}$$

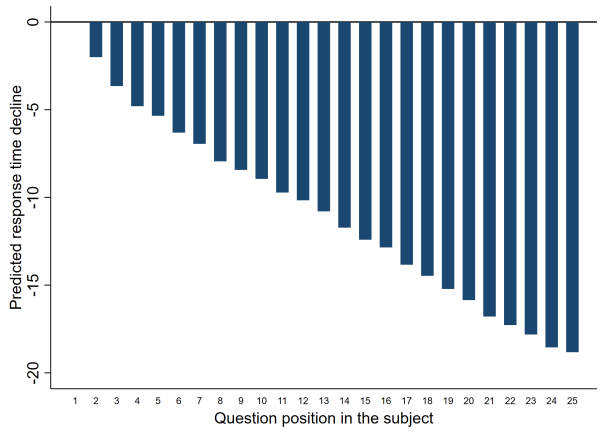
where $\widetilde{r}_{ij} = r_{ij} - \overline{r}_j$, $\widetilde{Q}^p_{ij} = Q^p_{ij} - \overline{Q}^p_j$, where $\overline{Q}^p_j$ is the average question position inside the respective block. Equation 2 is estimated both with and without student fixed effects. If the exclusion restriction is valid, the instrument should be independent of student fixed effects. A comparison between the two estimations is an indicative test for the validity of the IV approach.

Additionally, the instrument needs to be relevant, i.e. question position needs to be a strong predictor of response time. Figure 4 shows how response time develops across the test. The left-hand panel of figure 4 shows raw averages of response time over question position, while the right-hand panel shows estimates for question position dummies in a regression that also includes question fixed effects. Recall that there are different test versions for math and reading, and that higher tracks typically answer more, and longer questions. This explains the flattening out of the pattern at the end of the test in the raw graphs on the left. The corrected graphs, on the right, show that the decline is monotonic and roughly linear across subjects. Moreover, this pattern is persistently present also when split by gender or SES (see figure B1 in Appendix B.)

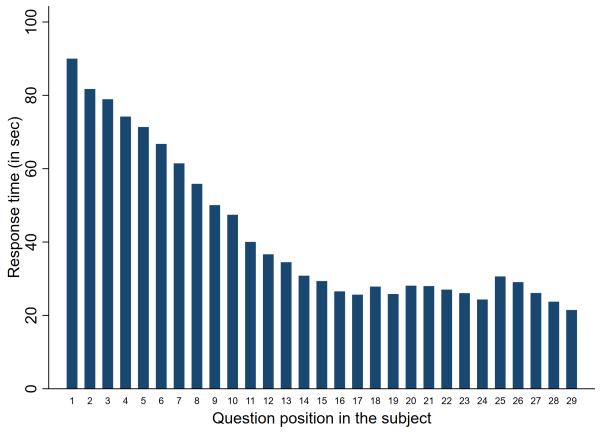
Figure 4: Response time over question position



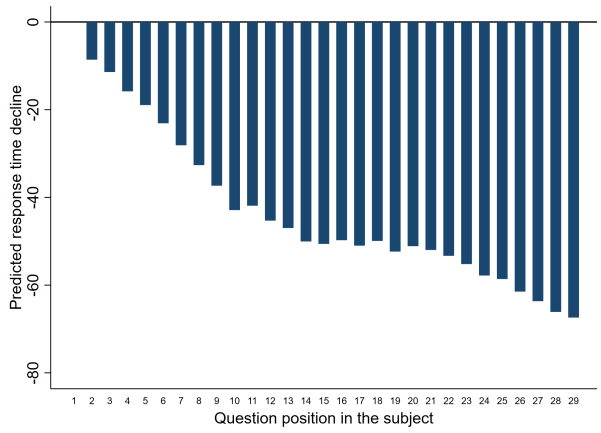
(a–1) IQ: absolute response time



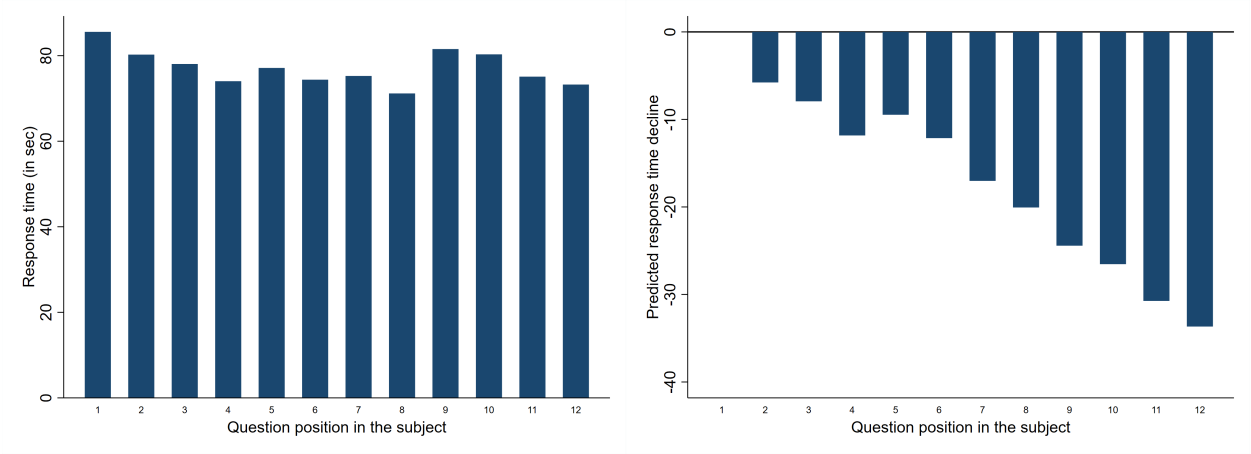
(a–2) IQ: predicted decline w.r.t 1st question



(b–1) Math:absolute response time



(b–2) Math: predicted decline w.r.t 1st question



(c–1) Reading: absolute response time

(c–2) Reading: predicted decline w.r.t 1st question

Note: Panels on the left-hand side of the figure show the average response time (in seconds) for each question position in the subject. Panels on the right-hand side of the figure show the predicted response time decline (in seconds) for each question position, when controlling for question fixed effects. This corresponds to the point estimates $\widehat{\psi}_k$ obtained from regressing response time on question fixed effects, subject question position and year dummies, namely $r_{ij} = \alpha_j + \sum_{k=1}^{K} \psi_k Q_k^p + \Psi Year_i' + w_{ij}$.

3.2 Estimating the implications for achievement gaps

Once the causal impact of response time on student performance is identified, we proceed to investigate whether differences in response time explain differences in performance, namely gender and SES achievement gaps. We first estimate raw achievement gaps for each subject, as shown in equation 3. These achievement gaps do not take the effect of response time on performance into account.

$$Raw\ Gap: \qquad score_{ij} = \alpha_j + \theta_1^R female_i + \theta_2^R high\ SES_i + \eta_1 V_i' + \epsilon_{ij}, \tag{3}$$

In equation 3, α_j are question fixed effects, $female_i$ takes value 1 for female students and 0 otherwise, $high\ SES_i$ takes value 1 when at least one of student i's parents has completed tertiary education and V_i' is a vector of school year dummies.[11] The raw achievement gaps are captured by the coefficient estimates $\widehat{\theta}_1^R$ and $\widehat{\theta}_2^R$ in equation 3, which correspond to the average difference in performance across gender and SES groups, respectively. Question fixed effects are included so that we compare achievement gaps for the same questions.

[11] Results remain the same if we control for other test features, such as test version and block.

We then estimate what remains of achievement gaps when differences in response time are controlled for. Consistent with the setup of Section 3.1, we show and compare this for different specifications. The naive OLS approach (equation 4) simply adds demeaned response time ($\tilde{r}_{ij}$) to the baseline regression (equation 3):

$$OLS\ Gap:\ score_{ij} = \alpha_j + \theta_1^{OLS} female_i + \theta_2^{OLS} high\ SES_i + \eta_2 V_i' + \beta_1 \tilde{r}_{ij} + \dot{\epsilon}_{ij} \tag{4}$$

Our preferred approach takes response time's endogeneity into account, by estimating the aforementioned IV approach that instruments response time with question position:

$$\begin{cases} \tilde{r}_{ij} = \alpha_j + \gamma_1 \widetilde{Q}_{ij}^p + \gamma_2 female_i + \gamma_3 high\ SES_i + \eta_3 V_i' + v_{ij} & \text{(5a)} \\ IV\ Gap:\ score_{ij} = \alpha_j + \beta_2 \hat{r}_{ij} + \theta_1^{IV} female_i + \theta_2^{IV} high\ SES_i + \eta_4 V_i' + \dot{v}_{ij} & \text{(5b)} \end{cases}$$

In model 5, $\widetilde{r}_{ij} = r_{ij} - \overline{r}_j$, $\widetilde{Q}_{ij}^p = Q_{ij}^p - \overline{Q}_j^p$. Note that $\hat{r}_{ij}$ are the fitted values from equation 5a.

Equations 4 and 5 correspond to the specifications without student fixed effects. fixed effects. To estimate achievement gaps including fixed effects we take a two-step procedure. First, we estimate student fixed effects, $\hat{\dot{\alpha}}_i$ from equation 1. Second, we regress student fixed effects on gender and SES. In essence, this extracts the gender and SES components from the individual fixed effects. This allows us to control for unobserved student characteristics and still ensure that the estimated effect of response time is statistically identical across the two regressions.

For the naive OLS approach, this means that student fixed effect are estimated from equation 1, and then regressed on gender and SES:

$$OLS\ Gap\ w/\ Student\ f.e.:\quad \hat{\dot{\alpha}}_i = \alpha_j + \theta_1^{OLSfe} female_i + \theta_2^{OLSfe} high\ SES_i + \eta_5 V_i' + \ddot{\epsilon}_{ij} \tag{6}$$

For the IV approach, student fixed effect are estimated from equation 1 but taking a 2SLS procedure that includes equation 2 as the first stage. Then we regress estimated student fixed effects on gender and SES:

$$IV\ Gap\ w/\ Student\ f.e.: \quad \hat{\ddot{\alpha}}_i = \alpha_j + \theta_1^{IVSfe} female_i + \theta_2^{IVSfe} + \eta_6 V_i' + \ddot{v}_{ij}, \tag{7}$$

We formally test whether these new point estimates are significantly different from the raw achievement gaps in equation 3, and between the different specifications that control for response time.[12]

Aside from how long students think about a question, they can also differ in how *hard* they think given a certain response time. Our implicit assumption is that a change in such *mental effort* along the test is proportional to gender and SES group differences in response time effort. We revisit this assumption in Section 4.3.

4 Results

This section is divided into two parts. In the first, we analyze the returns to response time. In the second, we evaluate to what extent achievement gaps change when response time is taken into account.

4.1 The causal effect of response time

Table 2 summarizes our findings for IQ, math and reading, respectively. The first three columns of each panel display OLS estimates. Fixed effects are included sequentially: column 1 does not include fixed effects, column 2 controls for question fixed effects and column 3 adds student fixed effects (equation 1).

The last column of table 2 shows the IV estimates of the second stage of our 2SLS (equations 1 and 2). Column 4 does the same but without student fixed effects. In the last rows of these columns, we report: (i) the first stage F-statistic, (ii) the p-value of a "Durbin-Wu-Hausman" test

[12] Testing whether achievement gaps change significantly from the base to the extended specifications is not straightforward. We use two different methods to perform such tests. One gives more accurate estimates, whereas the other allows for a more straightforward procedure to test our hypotheses. Further details may be found in Appendix D.

(DWH hereafter), and (iii) the F-statistic of the second stage regression. The first-stage F-statistics are far above conventional critical values, providing formal confirmation to the earlier visual result that response time declines significantly along the test.

Table 2: Returns to response time

	OLS			IV	
	(1)	(2)	(3)	(4)	(5)
Panel A: IQ					
Response time	0.051***	0.162***	-0.089***	0.007	0.011
	(0.010)	(0.011)	(0.010)	(0.055)	(0.055)
N	137,192	137,192	137,183	137,192	137,183
F-stat 1st stage				2737.414	2746.409
P-value DWH test				0.004	0.066
F-stat 2nd stage	18.542	82.638	84.926	8.928	0.039
Panel B: Math					
Response time	0.137***	0.243***	0.094***	0.356***	0.351***
	(0.003)	(0.003)	(0.003)	(0.018)	(0.018)
N	172,172	172,172	172,134	172,172	172,134
F-stat 1st stage				3927.763	3949.965
P-value DWH test				0.000	0.000
F-stat 2nd stage	893.726	2274.254	969.057	136.656	382.828
Panel C: Reading					
Response time	0.052***	0.163***	0.058***	0.283***	0.285***
	(0.002)	(0.003)	(0.003)	(0.025)	(0.024)
N	75,415	75,415	75,363	75,415	75,363
F-stat 1st stage				680.629	699.744
P-value DWH test				0.000	0.000
F-stat 2nd stage	191.579	1218.772	466.705	54.803	135.745
Question fixed effects	No	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	No	Yes

Notes: All specifications control for school year, unless student fixed-effects are included. Question score $\in [0, 100]$. Response time is demeaned and winsorized for right-tail outliers. Standard errors in parenthesis are clustered at student-level. Results remain significant if standard errors are clustered at both student and question levels. In the endogeneity test H_0 : *response time is exogenous*, with test statistic $\sim \chi^2_1$. See Baum *et al.* (2002) for details. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The DWH test allows us to verify whether OLS and IV estimates are equivalent. Its p-values reflect that test statistics are far above critical values, leading us to reject the null hypothesis that OLS and IV estimates are equal. Moreover, the low DWH p-values in column 5 confirm that an

instrumental variable is necessary even after including student and question fixed effects.[13]

For IQ, OLS estimates in the full specification (column 3 of table 2) are negative, whereas IV estimates are not statistically different from zero. Thus, regardless of how long one thinks about a question, the likelihood to provide a correct answer does not increase significantly with time invested. The negative OLS coefficient may be interpreted as reflecting the mismatch effect between question content and student knowledge. When the mismatch is higher, the probability to answer correctly is lower and students take longer to answer. When this channel is corrected, point estimates become statistically insignificant. Given the strong first stage power, this result is not driven by lack of a (meaningful) effort decline. The result may be due to the nature of what IQ tests measure. Fluid intelligence, typically dubbed as relation-perceiving ability (Cattell, 1987), may thus not be responsive to effort increases. Note that response times for IQ are also substantially shorter per question than for math and reading.

For math and reading, returns to response time are positive in all specifications. The results from the simple OLS model in column 1 confirm earlier observations: predicted performance increases with response time (Joseph, 2005). OLS estimates increase somewhat when question fixed effects are added, reflecting that students think longer on more difficult questions. They decrease again when student fixed effects are added, reflecting that higher-ability students think longer.

In the IV approach, the point estimates of response time remain positive and statistically significant, and effect sizes increase substantially. The comparison with the OLS estimates suggests that students answer worse and think longer on questions that they "personally" perceive as more difficult, leading to a downward bias in the estimated impact of response time on performance in OLS models. Column 5 of panels B and C of Table 2 suggests that a 10 seconds increase in average response time leads to an increase in the predicted probability of answering question j correctly of, approximately, 3.5 percentage points (pp hereafter) in math questions and 2.9pp in reading, all else equal. For math [reading] questions, this represents an 18.9 [12.8] percent increase in response

[13] Note that the specification in column 3 is the endogenous counterpart of column 5 used in the "DWH" test.

time and a 7.8 [4.3] percent increase in the predicted probability of answering a question correctly. Importantly, the IV models with and without student fixed effects provide near-identical results. This indicates that the instrument does not correlate with time-invariant unobserved student characteristics, providing evidence in favour of its validity. Compared to the most standard approach in the literature (OLS with student fixed effects; column (3)), IV estimates are around four times larger. The size of this bias is substantially larger than that caused by not controlling for students or question fixed effects.

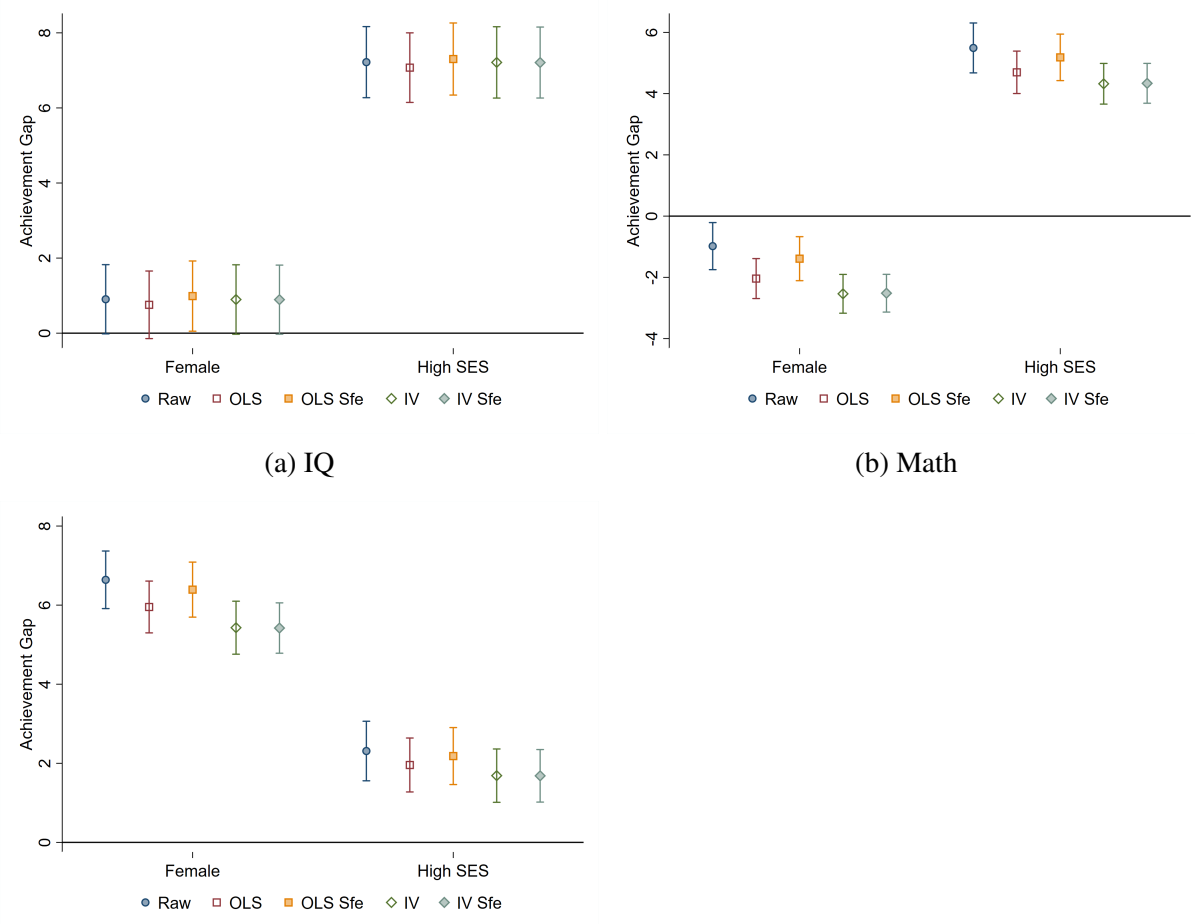
4.2 Implications for achievement gaps

We now evaluate to what extent gender and SES achievement gaps in IQ, math and reading change when response time is controlled for. We present results for all subjects, but our analysis focuses on math and reading. Firstly, question response time proved to be a relevant explanatory variable for performance in these subjects. Secondly, these are the achievement gaps that hold more academic and policy relevance. Results for IQ are depicted in the tables and figures for completeness.

Each panel of Figure 5 shows the estimated coefficients for female and high SES students, as indicated in the horizontal axis, for each subject. Each set of estimates is composed of five markers and their 95% confidence intervals. We start out with the raw gap (3), where response time is not controlled for. This is followed by the specifications in which response time is corrected for: the OLS and IV models, with and without student fixed effects. The underlying numbers for Figure 5 can be seen in Appendix Table C1.

Figure 5 shows that the gender achievement gap in reading decreases when response time is accounted for. This decrease is around 10 percent in the basic OLS approach, 4 percent in the OLS with student fixed effects, and 18 percent in the IV approach. The gender gap in math, which is in favour of boys, increases when accounting for response time. This increase is around 100 percent in the basic OLS model, 40 percent in the OLS with student fixed effects, and 150 percent in the IV model. While these numbers are sizable, it should be kept in mind that they are relative to an initial math gender gap that is relatively small. In an absolute sense, the change in the gender gap

Figure 5: Estimated achievement gaps across specifications



(a) IQ

(b) Math

(c) Reading

Notes: In the legend of each figure, the estimate labelled "Raw" stands for the raw achievement gap that does not control for response time. The estimate labelled "OLS" introduces this control, without correcting for its potential endogenous nature, whereas "OLS Sfe" extracts gender and SES achievement gaps from student fixed effects. The estimate labelled "IV" controls for the exogenous variation in response time, whereas "IV Sfe" extracts gender and SES achievement gaps from student fixed effects. All specifications control for question fixed effects and school year (unless student fixed effects are included). Regression tables may be found in Appendix C.

between the specifications is only slightly larger in math than in reading (1.5 versus 1.2; see Table C1).

With respect to SES achievement gaps, implications are very similar between math and reading. Both fall by around 15 percent in the basic OLS approach, 5 percent in the OLS with student fixed effects, and around 20 to 25 percent in the IV specifications. Taken together, these results indicate that the relevance of response time effort for achievement gaps is around 1.5 times larger than when using the OLS specification that only contains question fixed effects, and around 4 times larger than comparing to the standard approach of OLS with student fixed effects.

To sum up, effort differences explain around 21 percent of the SES gap in math achievement, 27 percent of the SES gap in reading achievement, and 18 percent of the gender gap in reading achievement. Conversely, accounting for effort increases the gender gap in math achievement by 157 percent.

While correcting for response time decreases the gender gap for one subject while increasing it for another, the correction favours the achievement for boys in both subjects. It is a well-established fact that girls have higher GPA's than boys, which likely is a major cause for their higher final educational attainment as well.[14] If we look at the gender gap in *average* achievement across math and reading, we find that response time explains around 49 percent.[15]

Lastly, we test whether these changes in achievement gaps across specifications are statistically different from zero. Comparing coefficient estimates across models is not straightforward, since a cross-model variance-covariance matrix must be estimated. We use two different methods to perform these tests, which lead to the same conclusion for both math and reading. Please see Appendix D for details. Appendix Table D4 summarizes our results, comparing achievement gaps from equation 7 to equations 3, 4 and 6 in each column, respectively.[16] At 5% significance level, we may conclude that the gender and SES achievement gaps change significantly for both math and reading.

4.3 Heterogeneity in the returns to response time

We have identified the return to response time using exogenous variation in question position, and used these estimates to infer how effort contributes to achievement gaps. This implicitly assumes that the causal effect of response time is homogeneous. Heterogeneous returns to response time may have implications for its contribution to achievement gaps (1) when there is nonlinearity in the returns to response time, and (2) when returns substantially differ by gender and/or SES.

[14] See, e.g., Breda & Napp (2019)

[15] Calculated by comparing the averaged raw gender gap across math and reading (column (1) in Table C1) to the averaged corrected gender gap across math and reading (column (5) in Table C1).

[16] Auxiliary regressions available in appendix D.1

Thirdly, given that our estimation approach exploits variation in response time along the test, our conclusions may also be affected if returns are different depending on the stage of the test. We review these three issues here, starting with the last.

Returns to response time may differ between beginning and end of the test. One could perceive effort as consisting of how *long* students think and how *hard* students think. The former (labeled D1) is the focus of our study which exploits that effort falls along the test; the latter (labeled D2) is difficult to measure directly but may also change along the test.

If our causal returns to response time also pick up on D2, this could lead to bias in how we use these estimates to explain achievement gaps. This bias depends on whether (level) differences in D2 by gender and SES are proportional to differences in D2 along the test. For example, if both D1 and D2 strongly drop along the test while effort differences by gender only consist of D1 differences, we are overestimating how strong achievement gaps would close when effort gaps would close.

We aim to identify whether D2 changes along the test, by examining the return to a second of thinking time along the test. If D2 falls along the test, returns to response time should be higher in the second half of the test (i.e. performance would drop stronger for a given drop in D1). We estimate this in the first and second panels of Table 3. In the first panel, we interact (demeaned) response time with a dummy that takes value 1 for the second half of the test. This corresponds to the student-specific empirical median of question position for each subject. In the second panel, we interact (demeaned) response time with question position in the subject. Estimates indicate that their is no evidence that returns differ along the test. This suggests that return estimates can be seen as clean from D2 changes, at least in this (relatively short) test. When we estimate returns using only the first half of the test and apply those results to the analysis on achievement gaps, results are virtually identical.

We conclude that there is little evidence that D2 changes along these tests and therefore our results in Figure 5 should be mainly interpreted as the effect on achievement gaps of closing the gap in D1. Moreover, under the assumption that gender/SES differences in D2 work in the same

direction as differences in D1, our estimates are lower bounds of the full contribution of effort to achievement gaps. Results from other literature suggest that D1 is relatively more malleable and therefore the more policy-relevant of the two dimensions. For instance, Borghans *et al.* (2008b) provide experimental evidence that financial incentives increase students' thinking time and performance simultaneously in a situation without time constraints, but that financial incentives do not increase student performance when time constraints are tight.

In panel B, we explore whether the relation between score and response is nonlinear. To do so, we add polynomial terms to our specification, namely the square of response time.[17] The point estimates for the squared term are statistically different from zero at 1% significance level. However, they are small in magnitude. The IV results in Section 4.1 (column 5 of panels B and C of table 2) suggest that an increase of 10 seconds in average response time leads to a percentage increase in the predicted probability of answering a question correctly of 7.8 [4.4] percent for math [reading] questions. When second-order polynomials are included, the corresponding figures are 8.8 and 5.2 percent, respectively. Thus, these approaches lead to qualitatively similar conclusions.

In panel C, we examine whether returns to response time differ by gender and SES groups, by conducting separate regressions for each group. We find no significant difference in response time returns across SES groups, but observe smaller returns for female students compared to males, particularly in reading questions. A 10-second increase in response time increases the predicted probability of correctly answering a reading question by 3.6 pp for males and 2.0 pp for females, compared to the average return of 2.8 pp that we identified before for the whole sample. The figures for math questions are more similar, with returns of 3.9 pp for males and 3.2 pp for females.

What does this heterogeneity imply for the relevance of effort for gender achievement gaps? Given the heterogeneous results in Panel C, it depends on how the effort gap is closed: are boys raising effort to the level of girls or are girls lowering effort to the level of boys (or a combination thereof). In the former case, the male return is relevant. A back-of-the-envelope calculation shows that this would imply that the gender achievement gap in reading closes by 25 percent. When using

[17] Appendix A provides details on our approach to instrumentalize response time squared.

Table 3: Heterogeneity in the returns to response time

	Math	Reading
Panel A		
$Response\ time_{ij}$	0.271***	0.271***
	(0.024)	(0.041)
$Response\ time_{ij} \times Half_{ij}$	0.034	0.019
	(0.051)	(0.065)
$Response\ time_{ij}$	0.417***	0.213***
	(0.037)	(0.079)
$Response\ time_{ij} \times Q_{ij}^{p}$	0.005	0.031
	(0.004)	(0.032)
Panel B		
$Response\ time_{ij}$	0.436***	0.348***
	(0.024)	(0.028)
$Response\ time_{ij}^{2}$	-0.004***	-0.001***
	(0.000)	(0.000)
Panel C		
$Response\ time_{ij}$	0.392***	0.355***
	(0.027)	(0.034)
$Response\ time_{ij} \times Female_{i}$	-0.076**	-0.151***
	(0.036)	(0.049)
$Response\ time_{ij}$	0.362***	0.283***
	(0.023)	(0.032)
$Response\ time_{ij} \times SES_{i}$	-0.030	0.008
	(0.036)	(0.050)

Notes: In panel A, the dummy variable $half_{ij}$ takes value 1 for the 2nd half of the test subject. This corresponds to questions taking place after the median question position of each subject. We use the empirical median for each student in each subject. Q_{ij}^{p} corresponds to question position in the subject. In panel C, we employ our 2SLS procedure and explore either the gender or the SES gap, corresponding to the first and second specification in panel C, respectively. Here, score is regressed on response time, an interaction term between the later and gender or SES (depending on the heterogeneity effect we are studying). All regressions control for student and question fixed effects and standard errors are clustered at the student level. In the first specification of panel A, fixed effects are interacted with the dummy variable $half_{ij}$; in the second specification, fixed effects are interacted with the variable Q_{ij}^{p}. In panel C, question fixed effects are interacted with either gender or SES, depending on the heterogeneity we are analysing.

the return estimates for girls (i.e. assuming that girls close the effort gap by reducing effort), the result is that the gap would close by 15 percent. The same calculation using the average return of 0.285 from Table 2 provides a figure of 20 percent. The relevant policy focus is likely the scenario in which boys raise effort, making our earlier results a modest underestimation of how effort contributes to the gender achievement gap in reading.

5 Conclusion

In this paper, we identify the causal return to response time and analyze how much of gender and SES achievement gaps can be explained by differences in test effort between these groups. We obtain causal estimates of the impact of question response time on student performance, relying on random question order. Our findings suggest that accounting for student effort closes the gender [SES] achievement gap in reading by 18 [27] percent, and the SES achievement gap in math by 21 percent. Gender achievement gaps in math widen by 157 percent when considered net of test effort. When we look at averaged achievement across both math and reading, effort can explain around 49 percent of the gender gap.

The main conclusion from the results of this study is not that closing these achievement gaps will now become easier. Whether closing effort differences is easy or hard is a direction for future research. Effective targeting of resources for remediation always depends on two aspects: importance and malleability. We have established the importance of effort in explaining achievement gaps, so future studies are needed to tackle the question of malleability. The results from this study have thus provided an alternative pathway through which achievement gaps may potentially be remediated. Additionally, it is not our message that the corrected achievement gaps are more accurate or better measures of gaps in educational achievement because they are net of effort. The ability to deliver effort, also in low-stakes tests, may be a valuable skill in itself, with payoffs in later life. Nonetheless, in order to effectively remediate achievement gaps, it is important that one understands whether they originate from a difference in ability/knowledge, or from a difference in

testing effort.

This study has provided a new perspective on achievement gaps by studying the (causal) role of effort. This new avenue provides opportunities for further extensions in future research. For one, the identified effects in this study pertain to a low-stakes test. The advantage of this setting is that it induces substantial variation in effort. Still, results may be different for high-stakes tests. For one, evidence shows that boys perform comparatively better on high-stakes tests compared to girls (Azmat *et al.*, 2016), which may be because their effort is more responsive to the change in stakes. If so, the relevance of effort for achievement gaps in high-stakes settings will be smaller. Nonetheless, it should be emphasized that low-stakes tests are extensively used in education, e.g., to assess retained knowledge, allowing teachers to adjust class content and pace. While doing so, it is relevant to take into account the effort students may put forth during these assessments. Additionally, low-stakes tests are highly dominant in research that aims to explain and target achievement gaps. We have shown that effort is relevant in these settings. Another interesting avenue for future studies would be to assess whether the typical strong variation in the gender gap in math across countries could be explained by country variation in gender effort gaps in these (also typically low-stakes) tests.

Secondly, more insights are needed on the underlying dimensions of test effort. While we have provided indicative evidence that our causal estimates of response time are not strongly driven by differences in how "hard" students think, the nature of response time differences still remains somewhat of a black box. Disentangling the components that underlie effort differences along the test and effort differences across groups may provide further insights into how effort can be targeted to remediate gaps in achievement.

Finally, the role of effort in explaining outcomes and inequalities can be carried outside of educational settings as well. Previous studies have shown that, e.g., careless answering behaviour in low-stakes tasks significantly correlates with labour outcomes (Zamarro *et al.*, 2018) and variation in performance decline along tests predicts economic growth (Balart *et al.*, 2018). In a similar way, one may use differences in response time, or decline in response time, in achievement tests to

predict outcomes in adult life. Alternatively, time investments made in non-educational settings, such as working life, and how these may explain differences in labour market success, provide an interesting avenue for future studies.

References

AKYOL, PELIN, KRISHNA, KALA, & WANG, JINWEN. 2021. Taking pisa seriously: How accurate are low-stakes exams? *Journal of labor research*, 1–60.

AZMAT, GHAZALA, CALSAMIGLIA, CATERINA, & IRIBERRI, NAGORE. 2016. Gender differences in response to big stakes. *Journal of the european economic association*, **14**(6), 1372–1400.

BALART, PAU, & OOSTERVEEN, MATTHIJS. 2019. Females show more sustained performance during test-taking than males. *Nature communications*, **10**(1).

BALART, PAU, OOSTERVEEN, MATTHIJS, & WEBBINK, DINAND. 2018. Test scores, noncognitive skills and economic growth. *Economics of education review*, **63**, 134–153.

BAUM, CHRISTOPHER F, SCHAFFER, MARK E, & STILLMAN, STEVEN. 2002. Ivreg2: Stata module for extended instrumental variables/2sls and gmm estimation. *Statistical software components, boston college department of economics*, 4.

BEDARD, KELLY, & CHO, INSOOK. 2010. Early gender test score gaps across oecd countries. *Economics of education review*, **29**(3), 348–363.

BHANOT, RUCHI, & JOVANOVIC, JASNA. 2005. Do parents' academic gender stereotypes influence whether they intrude on their children's homework? *Sex roles*, **52**(9), 597–607.

BORGHANS, LEX, & SCHILS, TRUDIE. 2018. Decomposing achievement test scores into measures of cognitive and noncognitive skills. *Available at ssrn 3414156*.

BORGHANS, LEX, DUCKWORTH, ANGELA LEE, HECKMAN, JAMES J, & TER WEEL, BAS. 2008a. The economics and psychology of personality traits. *Journal of human resources*, **43**(4), 972–1059.

BORGHANS, LEX, MEIJERS, HUUB, & WEEL, BAS TER. 2008b. The role of noncognitive skills in explaining cognitive test scores. *Economic inquiry*, **46**(1), 2–12.

BORGHANS, LEX, DIRIS, RON, & TAVARES, MARIANA. 2024. Student characteristics and effort during test-taking. *Learning and instruction*, **93**, 101924.

BORGONOVI, FRANCESCA, & BIECEK, PRZEMYSLAW. 2016. An international comparison of student ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and individual differences*, **49**(7), 128–137.

BREDA, THOMAS, & NAPP, CLOTILDE. 2019. Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the national academy of sciences*, **116**(31), 15435–15440.

BRUNELLO, GIORGIO, CREMA, ANGELA, & ROCCO, LORENZO. 2018. Testing at length if it is cognitive or non-cognitive. *Iza discussion paper no. 11603*.

CATTELL, RAYMOND BERNARD. 1987. *Intelligence: Its structure, growth and action*. Elsevier.

CECI, STEPHEN J, GINTHER, DONNA K, KAHN, SHULAMIT, & WILLIAMS, WENDY M. 2014. Women in academic science: A changing landscape. *Psychological science in the public interest*, **15**(3), 75–141.

COLEMAN, JAMES S. 1968. Equality of educational opportunity. *Integrated education*, **6**(5), 19–28.

CORNWELL, CHRISTOPHER, MUSTARD, DAVID B, & VAN PARYS, JESSICA. 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of human resources*, **48**(1), 236–264.

CORREIA, SERGIO. 2017. *Reghdfe: Stata module for linear and instrumental-variable/gmm regression absorbing multiple levels of fixed effects*. Tech. rept. Statistical Software Components s457874, Boston College Department of Economics.

CORREIA, SERGIO. 2018. Ivreghdfe: Stata module for extended instrumental variable regressions with multiple levels of fixed effects. *Statistical software components, boston college department of economics*, 9.

ELLISON, GLENN, & SWANSON, ASHLEY. 2023. Dynamics of the gender gap in high math achievement. *Journal of human resources*, **58**(5), 1679–1711.

FRYER JR, ROLAND G, & LEVITT, STEVEN D. 2010. An empirical analysis of the gender gap in mathematics. *American economic journal: Applied economics*, **2**(2), 210–40.

GNEEZY, URI, LIST, JOHN A., LIVINGSTON, JEFFREY A., QIN, XIANGDONG, SADOFF, SALLY, & XU, YANG. 2019. Measuring success in education: The role of effort on the test itself. *American economic review: Insights*, **1**(3), 291–308.

HANUSHEK, ERIC A, PETERSON, PAUL E, TALPEY, LAURA M, & WOESSMANN, LUDGER. 2019. *The unwavering ses achievement gap: Trends in us student performance*. Tech. rept. National Bureau of Economic Research.

HECKMAN, JAMES J. 2000. Policies to foster human capital. *Research in economics*, **54**(1), 3–56.

HERNÁNDEZ, MÓNICA, & HERSHAFF, JONATHAN. 2015. Skipping questions in school exams: The role of non-cognitive skills on educational outcomes. *Manuscript, univ. michigan. http://www. edpolicy. umich. edu/files/wp-hernandez-hershaff-skipping-questions-dec-2015. pdf.*

HIRSCH, STEFANIE. 2017. *Measurement in education: test scores and beyond*. Ph.D. thesis, Maastricht University.

JENNINGS, JENNIFER L, DEMING, DAVID, JENCKS, CHRISTOPHER, LOPUCH, MAYA, & SCHUELER, BETH E. 2015. Do differences in school quality matter more than we thought? new evidence on educational opportunity in the twenty-first century. *Sociology of education*, **88**(1), 56–82.

JOSEPH, E. 2005. Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology*, **125**, 88.

OECD. 2019. *PISA 2018 results (volume I): What students know and can do*. OECD. Chap. How much effort did students invest in the PISA test?, pages 198–207.

REARDON, SEAN F. 2016. School segregation and racial academic achievement gaps. *Rsf: The russell sage foundation journal of the social sciences*, **2**(5), 34–57.

SETZER, J. CARL, WISE, STEVEN L., VAN DEN HEUVEL, JILL R., & LING, GUANGMING. 2013. An investigation of examinee test-taking effort on a large-scale assessment. *Applied measurement in education*, **26**(1), 34–49.

SPENCER, STEVEN J, STEELE, CLAUDE M, & QUINN, DIANE M. 1999. Stereotype threat and women's math performance. *Journal of experimental social psychology*, **35**(1), 4–28.

STATACORP. 2015. Stata 15 base reference manual. *College station, tx: Stata press.*

STATACORP. 2017. Stata statistical software: Release 15. *College station, tx: Statacorp llc*.

TAVARES, MARIANA. 2022. *The magic number: Economic insights on achievement test scores*. Ph.D. thesis, Maastricht University.

WILDER, GITA Z, & POWELL, KRISTIN. 1989. Sex differences in test performance: A survey of literature. *College board report*.

WOOLDRIDGE, JEFFREY M. 2001. *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books, vol. 1, no. 0262232197. The MIT Press.

XIONG, XIAOLU, PARDOS, ZACHARY A., & T, NEIL. 2011. *An analysis of response time data for improving student performance prediction*.

ZAMARRO, GEMA, HITT, COLLIN, & MENDEZ, ILDEFONSO. 2016. When students don't care: Reexamining international differences in achievement and non-cognitive skills. *Edre working paper series*.

ZAMARRO, GEMA, CHENG, ALBERT, SHAKEEL, M. DANISH, & HITT, COLLIN. 2018. Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of behavioral and experimental economics*, **72**(2), 51–60.

A Including polynomials of response time

In this appendix we explain how we include higher order polynomials of response time in our IV regression with the goal of showing that our results are robust to a functional form that is not linear in this endogenous regression. The goal is to fit a 2SLS that includes a quadratic term of the endogenous regressor. To do so, we follow the advice in Section 9.5 of Wooldridge (2001). We first estimate the regression below and obtain its fitted values, $\hat{r}_{ij}$.

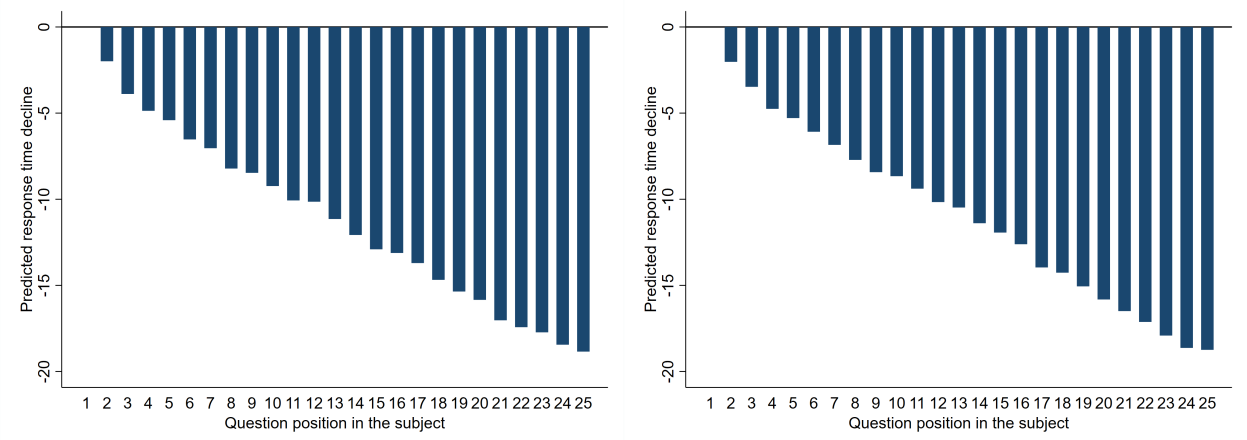
$$\tilde{r}_{ij} = \alpha_j + \gamma_0 \widetilde{Q}^p_{ij} + \upsilon_{ij}$$

We calculate the square of these fitted values, $\hat{r}^2_{ij}$, and use them as instruments for $\tilde{r}^2_{ij}$ (using Stata factor notation) in the regression below. This allows us to obtain exogenous estimates of β_{12}.

$$score_{ij} = \alpha_j + \ddot{\alpha}_i + \beta_{11}\tilde{r}_{ij} + \beta_{12}\tilde{r}^2_{ij} + u_{ij}$$

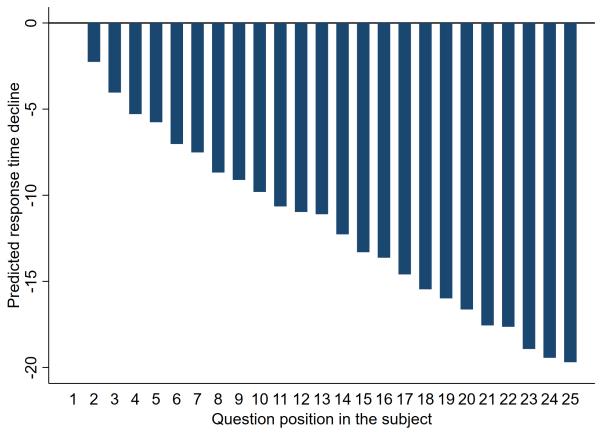
B Predicted response time decline across groups

Figure B1: Response time over question position

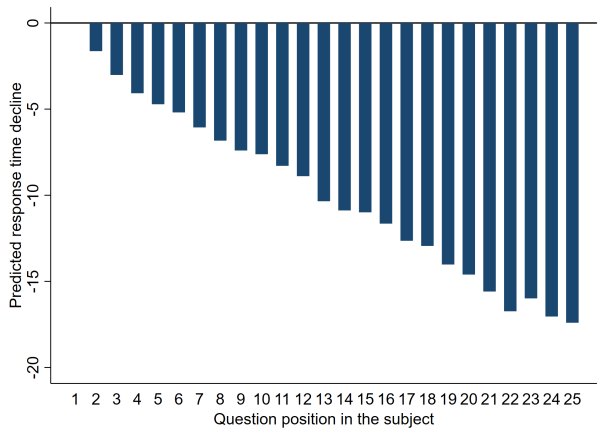


(a–1) IQ: males

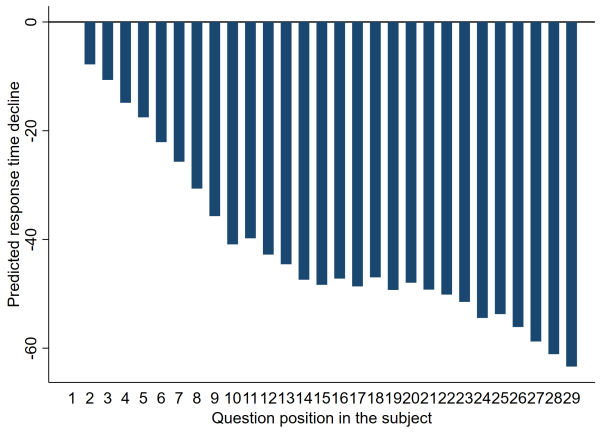
(a–2) IQ: females



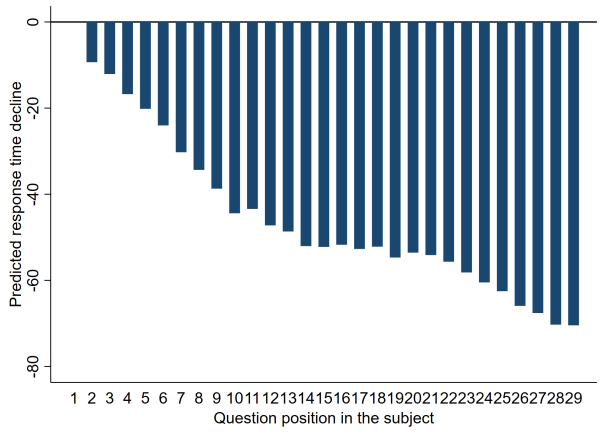
(b–1) IQ: Low SES



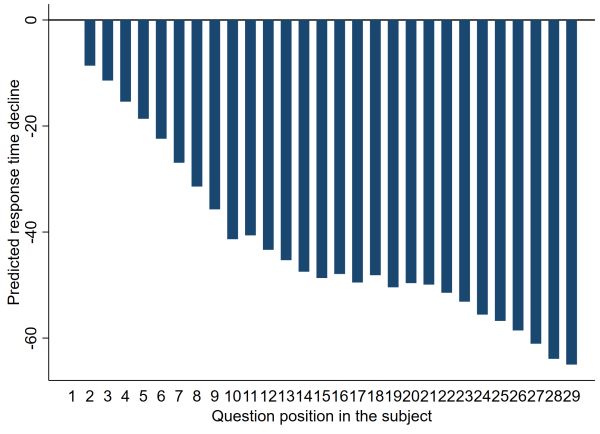
(b–2) IQ: High SES



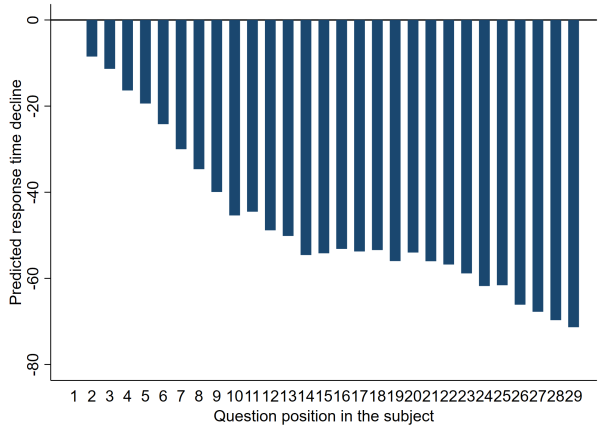
(c–1) Math: males



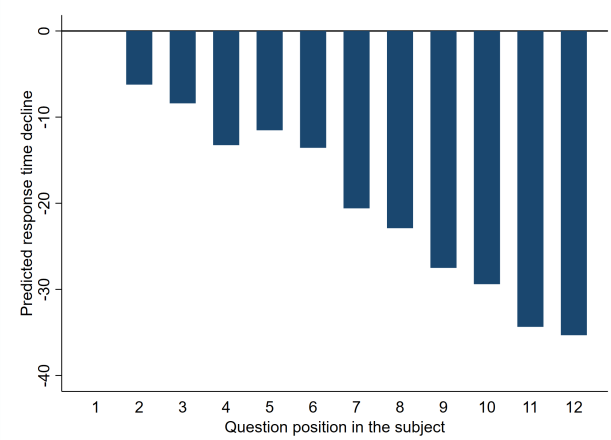
(c–2) Math: females



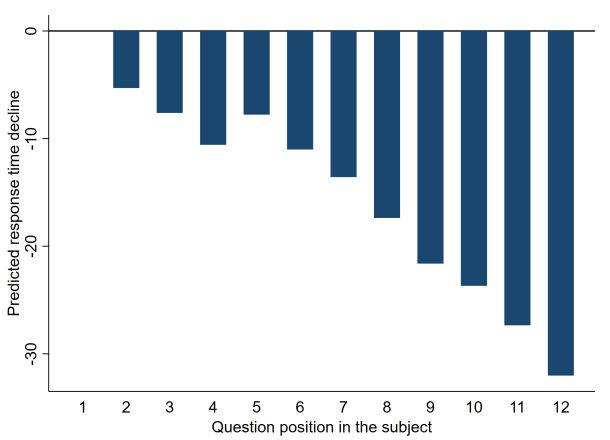
(d–1) Math: Low SES



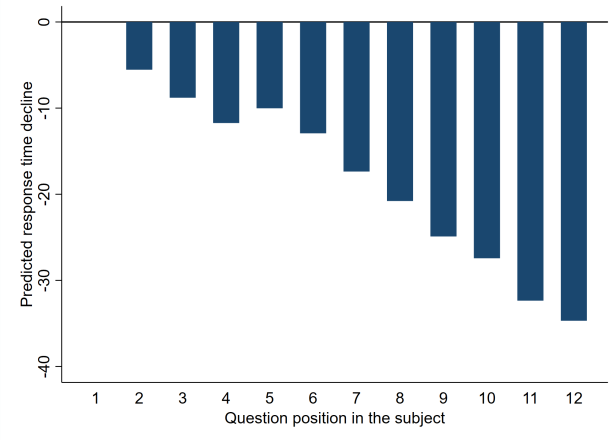
(d–2) Math: High SES



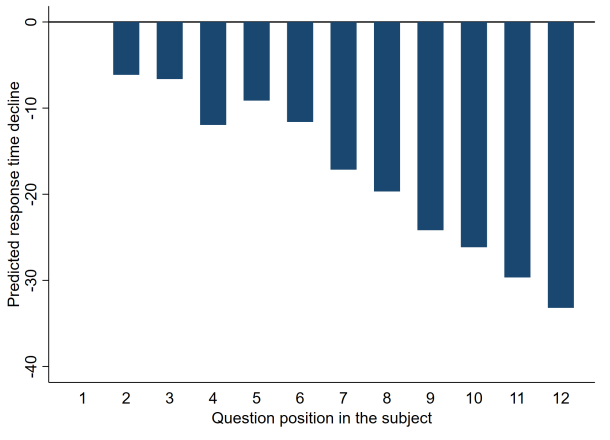
(e–1) Reading: males



(e–2) Reading: females



(f–1) Reading: Low SES



(f–2) Reading: High SES

Note: no estimate for 1st position as the decline is with respect to the 1st question position in the subject test.

C Estimated achievement gaps across specifications

This Appendix includes the regression tables associated with the point estimates presented in figure 5. All specifications control for academic year, unless student fixed-effects are included, and question fixed effects. Response time is demeaned and winsorized. Standard errors in parenthesis are clustered at student-level.

Table C1: Achievement Gaps

	(1) Raw	(2) OLS	(3) OLS Sfe	(4) IV	(5) IV Sfe
IQ questions					
Female	0.903^{+}	0.755^{+}	0.988*	0.897^{+}	1.194*
	(0.470)	(0.459)	(0.477)	(0.472)	(0.529)
High SES	7.220***	7.073***	7.303***	7.213***	7.272***
	(0.483)	(0.473)	(0.490)	(0.485)	(0.546)
N	137,192	137,192	137,183	137,192	137,183
F-stat	64.155	97.283	63.795	51.444	51.434
Math questions					
Female	-0.980*	-2.040***	-1.390***	-2.537***	-2.518***
	(0.392)	(0.333)	(0.366)	(0.323)	(0.315)
High SES	5.492***	4.696***	5.184***	4.322***	4.337***
	(0.416)	(0.353)	(0.387)	(0.339)	(0.332)
N	172,172	172,172	172,134	172,172	172,134
F-stat	49.872	1448.670	52.058	141.426	60.764
Reading questions					
Female	6.640***	5.953***	6.392***	5.429***	5.420***
	(0.371)	(0.334)	(0.355)	(0.342)	(0.325)
High SES	2.310***	1.957***	2.183***	1.688***	1.683***
	(0.384)	(0.348)	(0.367)	(0.344)	(0.339)
N	75,415	75,415	75,363	75,415	75,363
F-stat	94.496	799.874	95.837	124.268	82.435
IV	No	No	No	Yes	Yes
Response time	No	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	No	Yes

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

D Estimation and Testing

This appendix describes how test hypotheses are conducted across models. Performing such tests allows us to verify whether achievement gaps change significantly across specifications. We use two estimation methods, conferring robustness to our results. Each method has advantages and disadvantages, explored below.

This appendix is organized in subsections according to the estimation method employed. The first subsection describes our preferred estimation technique. It relies on a user-written Stata command that increases the precision of IV estimates but does not allow for straightforward testing across specifications. The second subsection uses a literature standard estimation procedure that allows for straightforward testing across specifications, but whose predictions display higher standard errors.

D.1 Method 1: Using *ivreghdfe*

Given the fractional nature of our dependent variable, we would be interested in a command or method that estimates a nonlinear model, e.g.: probit or logit. On the other hand, we must deal with potential endogeneity of response time, a large set of fixed effects, and unobserved factors at the individual level that require clustered standard error. At the time of our analysis, and to the best of our knowledge, the most accurate, efficient and time-saving method to deal with such data and model features, in Stata 17 (StataCorp, 2017), is the user-written command *ivreghdfe* (Baum *et al.*, 2002; Correia, 2017, 2018). It allows for:

1. high dimensional fixed effects;
2. the use of instrumental variables to correct for endogeneity;
3. clustering of standard errors at a level different than the fixed effect;
4. panels that are not nested within clusters.

In this sense, our best alternative is to estimate a linear model using *ivreghdfe*, overcoming estimation and feasibility limitations of nonlinear models. To be consistent, we employ *reghdfe* (Correia, 2017) to estimate specifications that do not suffer from endogeneity, as is the case of equation 3.

Although *ivreghdfe* allows us to obtain consistent and efficient estimates for achievement gaps, it does not allow for a straightforward method to compare point estimates across specifications. This is relevant for our analysis as we are interested in assessing whether achievement gaps change significantly when response time is included in the model.

To perform such a test it is necessary to estimate the covariance between point estimates across specifications. To test the equality of two coefficients, $H_0 : \theta = \delta$, the appropriate test statistic is:

$$t = \frac{\hat{\theta} - \hat{\delta}}{\hat{\sigma}_{(\hat{\theta}-\hat{\delta})}} = \frac{\hat{\theta} - \hat{\delta}}{\sqrt{\hat{\sigma}^2_{\hat{\theta}} + \hat{\sigma}^2_{\hat{\delta}} - 2 \times \hat{\sigma}_{\hat{\theta},\hat{\delta}}}} \sim \chi^2_1$$

This may be obtained by estimating a seemingly unrelated regression (SUR model), using the Stata command *suest* (StataCorp, 2015). However the latter is incompatible with *ivreghdfe*, relying on OLS regressions (*reg*). In order to circumvent this problem, we mimic *suest*'s procedure and adapt it to our case. Our approach finds inspiration in example 3 of StataCorp (2015), under the title *SUEST*.

We compare the estimated achievement gap from equation 7, which corrects for response time edonegeneity and unobserved student characteristics, with the raw achievement gap from equation 3, the OLS gaps from equations 4 and 6. The latter also takes into account student unobserved characteristics. All estimated achievement gaps take into account question characteristics. To perform such comparisons, we stack our data as explained below, run auxiliary regressions and test whether the relevant coefficients are statistically equal to 0.

Matrix 8 shows how data is stacked to compare estimated achievement gap from equation 7 to the ones estimated in equation 3 for a fixed number of students, N. We create a dummy variable that identifies each specification ($model$) and a dependent variable, y_{ij}, that for equation 3 ($model = 0$) corresponds to student's i score in question j ($score_{ij}$) and for equation 7 ($model =$

1) corresponds to student's i fixed effect from equation 2 ($\widehat{\ddot{\alpha}}_i$).

$$
\begin{array}{|cccccccc|}
\hline
model & studentid(i) & question(j) & y_{ij} & fem_i & ses_i & year_i \\
\hline
0 & 1 & A & score_{1,A} & fem_1 & ses_1 & year_1 \\
0 & 1 & B & score_{1,B} & fem_1 & ses_1 & year_1 \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
0 & N & Z & score_{N,Z} & fem_N & ses_N & year_N \\
1 & 1 & A2 & \widehat{\ddot{\alpha}}_1 & fem_1 & ses_1 & year_1 \\
1 & 1 & B2 & \widehat{\ddot{\alpha}}_1 & fem_1 & ses_1 & year_1 \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
1 & N & Z2 & \widehat{\ddot{\alpha}}_N & fem_N & ses_N & year_N \\
\hline
\end{array}
\tag{8}
$$

The next step is to estimate the following equation, with clustered standard errors at the student-level, using *reghdfe*, as we do throughout our analysis. Please note that we estimate each subject separately.

$$
\begin{aligned}
y_{ij} = \alpha_j + \lambda_1 female_i + \lambda_2 high\ SES_i + \lambda_3 year_i + \lambda_4 model + \\
+ \lambda_5 female_i \times model + \lambda_6 high\ SES_i \times model + \lambda_7 year_i \times model + \eta_{ij}
\end{aligned}
$$

Coefficient estimates for λ_1 and λ_2, along with their standard errors, match their counterparts obtained for equation 3 in our main analysis. Moreover, the sum of each main term with its associated interaction term matches its estimated counterpart for equation 7.

Testing whether θ_1^R is statistically different from θ_1^{IVSfe} is equivalent to test if λ_5 is different from 0. In the same manner, $H_0 : \theta_2^R = \theta_2^{IVSfe}$ is equivalent to $H_0 : \lambda_6 = 0$. These λ estimates correspond to the difference between estimated achievement gaps across equations 3 and 7. When

employing *reghdfe*, λ_5 and λ_6 capture the change in gender and SES achievement gaps, respectively, stemming from the introduction of (demeaned) exogenous response time. In this manner, we are left with a standard, within specification, t-test that is automatically reported by Stata after any estimation command, namely $H_0 : \lambda_k = 0$ for $k = 5, 6$. Results are shown in column 3 of tables D1, D2 and D3 for IQ, math and reading, respectively.

We are also interested in comparing estimates of equation 4 with the ones from equation 7. For this purpose, we re-arrange our data in the following manner:

$$\begin{bmatrix} model & studentid(i) & question(j) & y_{ij} & \ddot{r}_{ij} & fem_i & ses_i & year_i \\ 1 & 1 & A & score_{1,A} & \tilde{r}_{1,A} & fem_1 & ses_1 & year_1 \\ 1 & 1 & B & score_{1,B} & \tilde{r}_{1,B} & fem_1 & ses_1 & year_1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & N & Z & score_{N,Z} & \tilde{r}_{N,Z} & fem_N & ses_N & year_N \\ 0 & 1 & A2 & \widehat{\ddot{\alpha}}_1 & 0 & fem_1 & ses_1 & year_1 \\ 0 & 1 & B2 & \widehat{\ddot{\alpha}}_1 & 0 & fem_1 & ses_1 & year_1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & N & Z2 & \widehat{\ddot{\alpha}}_N & 0 & fem_N & ses_N & year_N \end{bmatrix} \tag{9}$$

We add a column with response time and reverse the model dummy. The next step is to estimate the following specification using *reghdfe* and clustering standard errors at the student-level:

$$y_{ij} = \alpha_j + \varphi_1 female_i + \varphi_2 high\ SES_i + \varphi_3 year_i + \varphi_4 \ddot{r}_{ij} + \varphi_5 model +$$
$$+ \varphi_6 female_i \times model + \varphi_7 high\ SES_i \times model + \varphi_8 year_i \times model + \varphi_9 \ddot{r}_{ij} \times model + \zeta_{ij}$$

In this case, estimates of φ_1, φ_2, φ_3 and φ_4, along with their standard errors, match the ones of equation 7. In a similar fashion as before, $\varphi_1 + \varphi_6$, $\varphi_2 + \varphi_7$, $\varphi_3 + \varphi_8$ and $\varphi_4 + \varphi_9$, as well as their standard errors, match the coefficient estimates from equation 4. Our interest lies on φ_6 and φ_7 and whether these are statistically different from zero. As before, a simple test (or evaluation of the estimates' p-values) suffices. Results are shown in column 5 of tables D1, D2 and D3 for IQ, math and reading respectively.

Lastly, to compare OLS estimates from equation 6) with IV estimates from equation 7, we organize our data as shown in matrix 10. The dependent variable (y_{ij}) now stacks student fixed from equation 1 ($\widehat{\ddot{\alpha}}_i^{OLS}$) and from equation 2 ($\widehat{\ddot{\alpha}}_1^{IV}$). Results are shown in column 7 of tables D1, D2 and D3 for IQ, math and reading respectively.

$$
\begin{bmatrix}
model & studentid(i) & question(j) & y_{ij} & fem_i & ses_i & year_i \\
\hline
0 & 1 & A & \widehat{\ddot{\alpha}}_1^{OLS} & fem_1 & ses_1 & year_1 \\
0 & 1 & B & \widehat{\ddot{\alpha}}_1^{OLS} & fem_1 & ses_1 & year_1 \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
0 & N & Z & \widehat{\ddot{\alpha}}_N^{OLS} & fem_N & ses_N & year_N \\
1 & 1 & A2 & \widehat{\ddot{\alpha}}_1^{IV} & fem_1 & ses_1 & year_1 \\
1 & 1 & B2 & \widehat{\ddot{\alpha}}_1^{IV} & fem_1 & ses_1 & year_1 \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
. & . & . & . & . & . & . \\
1 & N & Z2 & \widehat{\ddot{\alpha}}_N^{IV} & fem_N & ses_N & year_N
\end{bmatrix}
\tag{10}
$$

Table D1: Auxiliary regressions to test $H_0 : \Delta gap = 0$ for IQ questions

	(1) IV Sfe	(2) Raw	(3) Aux 1	(4) OLS	(5) Aux 2	(6) OLS Sfe	(7) Aux 3
Female	0.893^{+}	0.903^{+}	0.903^{+}	0.755^{+}	0.893^{+}	0.988*	0.988*
	(0.469)	(0.470)	(0.470)	(0.459)	(0.469)	(0.477)	(0.477)
Female $\times$ 1.model			-0.010***		-0.138***		-0.095***
			(0.002)		(0.029)		(0.018)
High SES	7.209***	7.220***	7.220***	7.073***	7.209***	7.303***	7.303***
	(0.482)	(0.483)	(0.483)	(0.473)	(0.482)	(0.490)	(0.490)
High SES $\times$ 1.model			-0.010***		-0.136***		-0.093***
			(0.002)		(0.029)		(0.019)
N	137,183	137,192	274,347	137,192	274,347	137,183	274,366
F-stat	64.197	64.155	34.864	97.283	59.396	63.795	36.360

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D2: Auxiliary regressions to test $H_0 : \Delta gap = 0$ for Math questions

	(1) IV Sfe	(2) Raw	(3) Aux 1	(4) OLS	(5) Aux 2	(6) OLS Sfe	(7) Aux 3
Female	-2.518***	-0.980*	-0.980*	-2.040***	-2.518***	-1.390***	-1.390***
	(0.315)	(0.392)	(0.392)	(0.333)	(0.315)	(0.366)	(0.366)
Female × 1.model			-1.538***		0.478***		-1.128***
			(0.157)		(0.051)		(0.115)
High SES	4.337***	5.492***	5.492***	4.696***	4.337***	5.184***	5.184***
	(0.332)	(0.416)	(0.416)	(0.353)	(0.332)	(0.387)	(0.387)
High SES × 1.model			-1.156***		0.359***		-0.847***
			(0.170)		(0.061)		(0.123)
N	172,134	172,172	344,306	172,172	344,306	172,134	344,344
F-stat	60.764	49.872	63.182	1448.670	1165.953	52.058	63.186

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D3: Auxiliary regressions to test $H_0 : \Delta gap = 0$ for Reading questions

	(1) IV Sfe	(2) Raw	(3) Aux 1	(4) OLS	(5) Aux 2	(6) OLS Sfe	(7) Aux 3
Female	5.420***	6.640***	6.640***	5.953***	5.420***	6.392***	6.392***
	(0.325)	(0.371)	(0.371)	(0.334)	(0.325)	(0.355)	(0.355)
Female × 1.model			-1.220***		0.533***		-0.972***
			(0.181)		(0.080)		(0.144)
High SES	1.683***	2.310***	2.310***	1.957***	1.683***	2.183***	2.183***
	(0.339)	(0.384)	(0.384)	(0.348)	(0.339)	(0.367)	(0.367)
High SES × 1.model			-0.627***		0.274**		-0.500***
			(0.187)		(0.083)		(0.148)
N	75,363	75,415	150.778	75,415	150.778	75,363	150,726
F-stat	82.435	94.496	48.581	799.874	467.320	95.837	48.610

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D4: Do achievement gaps change significantly? Using *ivreghdfe*

	Gender			SES		
IV w/ Sfe *vs*:	Raw	OLS	OLS w/ Sfe	Raw	OLS	OLS w/ Sfe
IQ	0.000	0.000	0.000	0.000	0.000	0.000
Math	0.000	0.000	0.000	0.000	0.000	0.000
Reading	0.000	0.000	0.000	0.001	0.001	0.001

The table shows p-values corresponding to $H_0 : \lambda_{increment} = 0$, as explained in Appendix D.1. Rejecting the null hypothesis suggests that achievement gaps change significantly from the specification in the column title compared to the one estimated in equation 7. For instance, the first column labelled "Raw" shows the p-value corresponding to testing whether the estimated raw achievement gap equals the one estimated in equation 7. As p-values are very close to zero, we reject the null hypothesis for all usual significance levels, and conclude that achievement gaps change significantly.

D.2 Method 2: Using *SUEST*

To further validate the approach taken in the subsection above, we run a set of OLS regressions, using the Stata command *reg* (StataCorp, 2015), and subsequently use *suest* to test the hypotheses across specifications.[18] In order to add question fixed effects, we make use of Stata factor notation. Standard errors are then clustered when running *suest*. For the 2SLS estimation, we start by obtaining predictions for the first stage (equation 5a) in a similar fashion as the one explained above. We include them in the second stage which is also estimated by means of *reg* and factor variable notation to account for question fixed effects.

Table D5 shows p-values for the method presented in D.2. It relies on estimating seemingly unrelated regressions and the covariance between tested estimates. Rejecting the null hypothesis suggests that the change in point estimates is statistically different from zero. Main results remain unchanged regardless of the method used.

[18] Stata commands as *ivreg*, *areg*, *xtreg* (StataCorp, 2015) are not supported by *suest*. Thus, we are left with *reg*.

Table D5: Do achievement gaps change significantly? Using *SUR*

	Gender			SES		
IV w/ Sfe *vs*:	Raw	OLS	OLS w/ Sfe	Raw	OLS	OLS w/ Sfe
IQ	0.000	0.000	0.000	0.000	0.000	0.000
Math	0.000	0.000	0.000	0.000	0.000	0.000
Reading	0.000	0.000	0.000	0.0008	0.001	0.0007

The table shows p-values for the equality of achievement gaps from a Seemingly Unrelated Estimation, as explained in Appendix D.2. Rejecting the null hypothesis suggests that achievement gaps change significantly from the specification in the column title compared to the one estimated in equation 7. For instance, the first column labelled "Raw" shows the p-value corresponding to testing whether the estimated raw achievement gap equals the one estimated in equation 7. As p-values are very close to zero, we reject the null hypothesis for all usual significance levels, and conclude that achievement gaps change significantly.

The tables below allows to compare the estimated achievement gaps for each approach explained in this appendix. For math and reading especially, standard errors differ between the two approaches in the IV estimation (columns 3 and 6). Point estimates are less precise when using *reg* compared to *ivreghdfe*, leading to lower t-statistics and higher p-values, which make it harder to reject the null hypothesis. This may explain the different conclusions regarding the SES gap in reading, for instance (see table D4).

Table D6: Achievment Gaps – IQ questions

	Using *ivreghdfe*				Using SUR (based on *suest and reg*)		
	(1) Raw	(2) OLS	(3) OLS Sfe	(4) IV Sfe	(5) Raw vs IV Sfe	(6) OLS vs IV Sfe	(7) OLS Sfe vs IV Sfe
Female	0.903^{+}	0.755^{+}	0.988*	0.893^{+}	0.903^{+}	0.755^{+}	0.988*
	(0.470)	(0.459)	(0.477)	(0.469)	(0.470)	(0.459)	(0.477)
High SES	7.220***	7.073***	7.303***	7.209***	7.220***	7.073***	7.303***
	(0.483)	(0.473)	(0.490)	(0.482)	(0.483)	(0.473)	(0.490)
response time		0.156***				0.156***	
		(0.011)				(0.011)	
IV Sfe: mean							
Female					0.893^{+}	0.893^{+}	0.893^{+}
					(0.469)	(0.469)	(0.469)
High SES					7.209***	7.209***	7.209***
					(0.482)	(0.482)	(0.482)
N	137,192	137,192	137,183	137,183	137,183	137,183	137,183
F-stat	64.155	97.283	63.795	51.444	64.197		

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D7: Achievment Gaps – Math questions

	Using *ivreghdfe*				Using SUR (based on *suest and reg*)		
	(1) Raw	(2) OLS	(3) OLS Sfe	(4) IV Sfe	(5) Raw vs IV Sfe	(6) OLS vs IV Sfe	(7) OLS Sfe vs IV Sfe
main							
Female	-0.980*	-2.040***	-1.390***	-2.518***	-0.980*	-2.040***	-1.390***
	(0.392)	(0.333)	(0.366)	(0.315)	(0.392)	(0.333)	(0.366)
High SES	5.492***	4.696***	5.184***	4.337***	5.492***	4.696***	5.184***
	(0.416)	(0.353)	(0.387)	(0.332)	(0.416)	(0.353)	(0.387)
response time		0.242***				0.242***	
		(0.003)				(0.003)	
IV Sfe: mean							
Female					-2.518***	-2.518***	-2.518***
					(0.314)	(0.314)	(0.314)
High SES					4.337***	4.337***	4.337***
					(0.332)	(0.332)	(0.332)
N	172,172	172,172	172,134	172,134	172,134	172,134	172,134
F-stat	49.872	1448.670	52.058	60.764			

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D8: Achievment Gaps – Reading questions

	Using *ivreghdfe*				Using SUR (based on *suest and reg*)		
	(1) Raw	(2) OLS	(3) OLS Sfe	(4) IV Sfe	(5) Raw vs IV Sfe	(6) OLS vs IV Sfe	(7) OLS Sfe vs IV Sfe
main							
Female	6.640***	5.953***	6.392***	5.420***	6.640***	5.953***	6.392***
	(0.371)	(0.334)	(0.355)	(0.325)	(0.371)	(0.334)	(0.355)
High SES	2.310***	1.957***	2.183***	1.683***	2.310***	1.957***	2.183***
	(0.384)	(0.348)	(0.367)	(0.339)	(0.384)	(0.348)	(0.367)
response time		0.161***				0.161***	
		(0.003)				(0.003)	
IV Sfe: mean							
Female					5.420***	5.420***	5.420***
					(0.325)	(0.325)	(0.325)
High SES					1.683***	1.683***	1.683***
					(0.339)	(0.339)	(0.339)
N	75,415	75,415	75,363	75,363	75,363	75,363	75,363
F-stat	94.496	799.874	95.837	82.435			

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

E Additional tables

Table E1: Randomization test for question order

Subject – block	P-values of Pearsonχ^2 test		
IQ	0.648 (24)		
	test version 1	*test version 2*	*test version 3*
Math – block A	0.99 (8)	0.85 (9)	0.59 (9)
Math – block B	1.00 (6)	1.00 (9)	1.00 (13)
Math – block C	1.00 (7)	1.00 (7)	0.99 (4)
Reading – block A	0.99 (3)	0.99 (3)	0.99 (5)
Reading – block B	1.00 (3)	0.99 (3)	1.00 (5)

P-values for Pearsonχ^2 tests with H_0 : *question order is random* for each block of each subject. Degrees of freedom in parenthesis.