Institute of Education

UCL

Do teenage boys perform less well than teenage girls in literacy or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC

PIAAC International Conference. Rome 28-29 2020

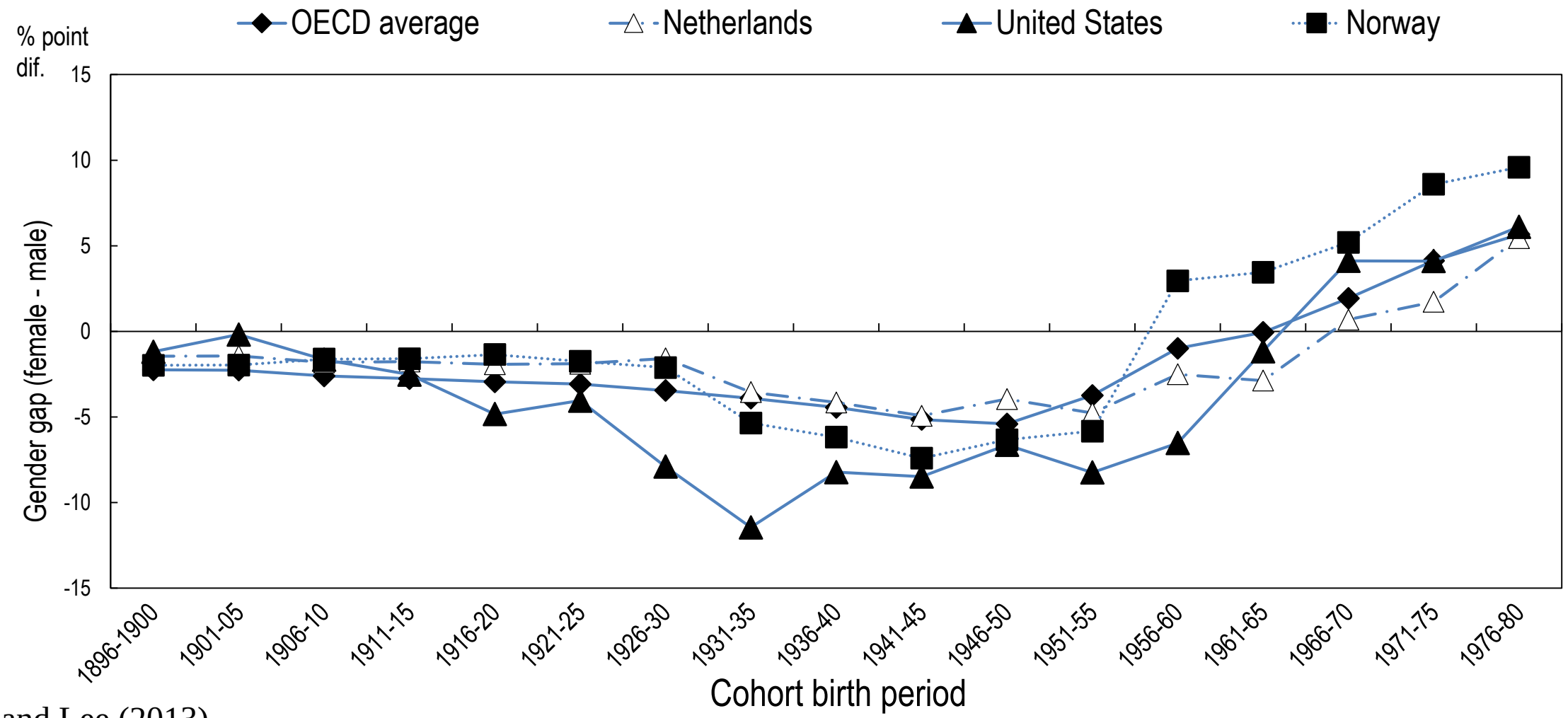
Francesca Borgonovi
British Academy Global Professor
Institute of Education – University College London

IOE LONDON

A ***VERY*** brief recap of relevant stylized facts on gender gaps

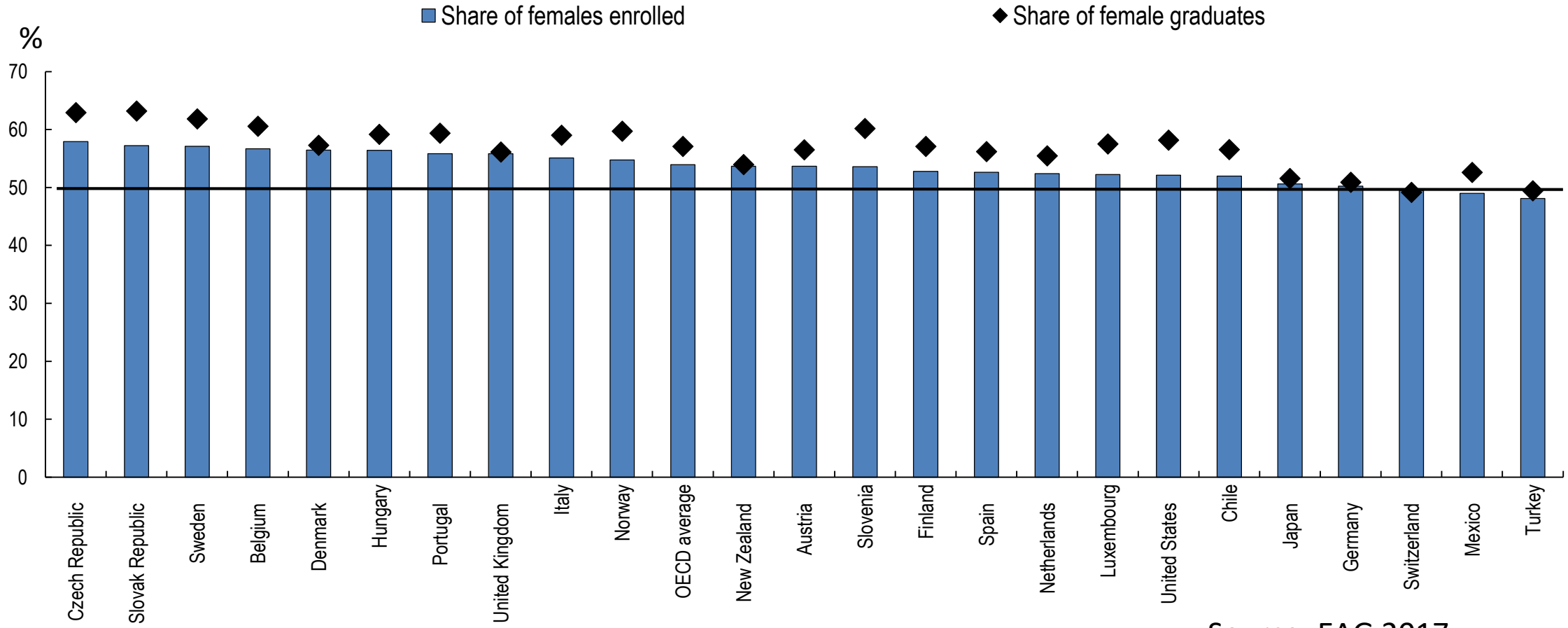
- Young females on average outperform young males in educational outcomes
  - Lack of motivation,
  - Higher rates of reading disorders,
  - Delayed cognitive development,
  - Dip in cognitive abilities and executive control during the teenage years
- Males continue to do better than females in the labour market
  - Discrimination,
  - Differences in division of labour within the family,
  - Differences in study and career choices

Long term trends in the gender gap in tertiary attainment



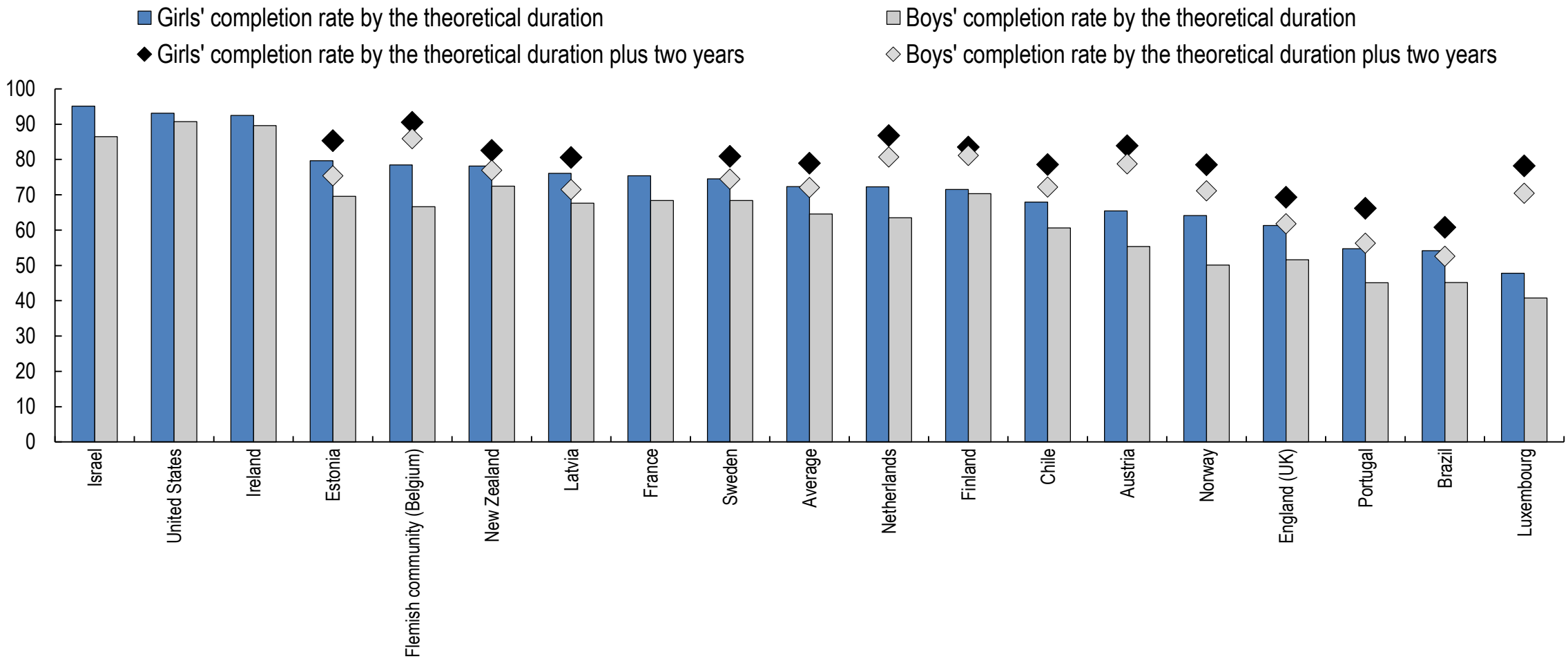
Barro and Lee (2013)

Gender gaps in tertiary education enrollment and completion



Source: EAG 2017.

Completion rate of upper secondary education by gender (2015)



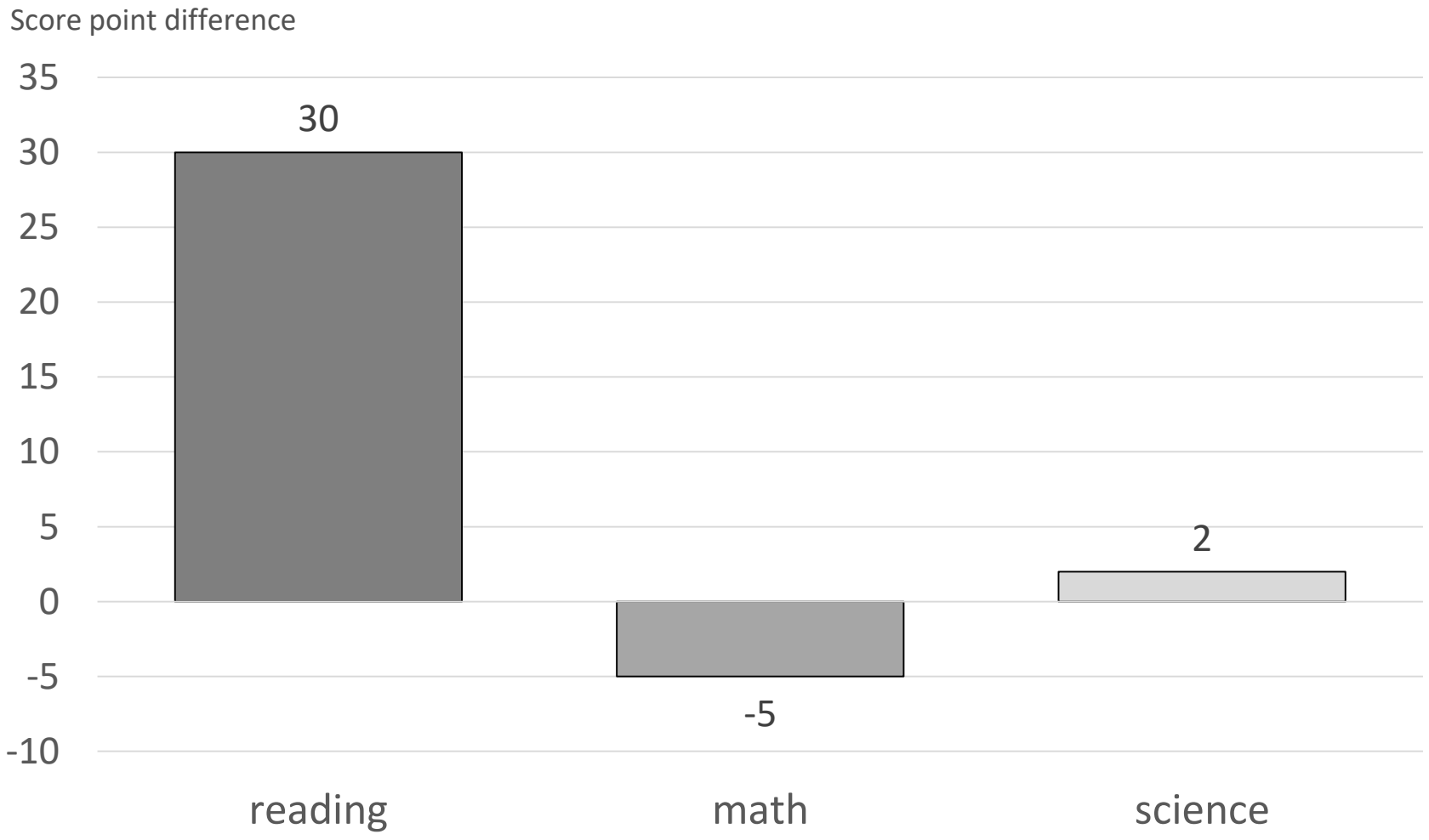
Source: EAG 2017.

The role of international large-scale assessments

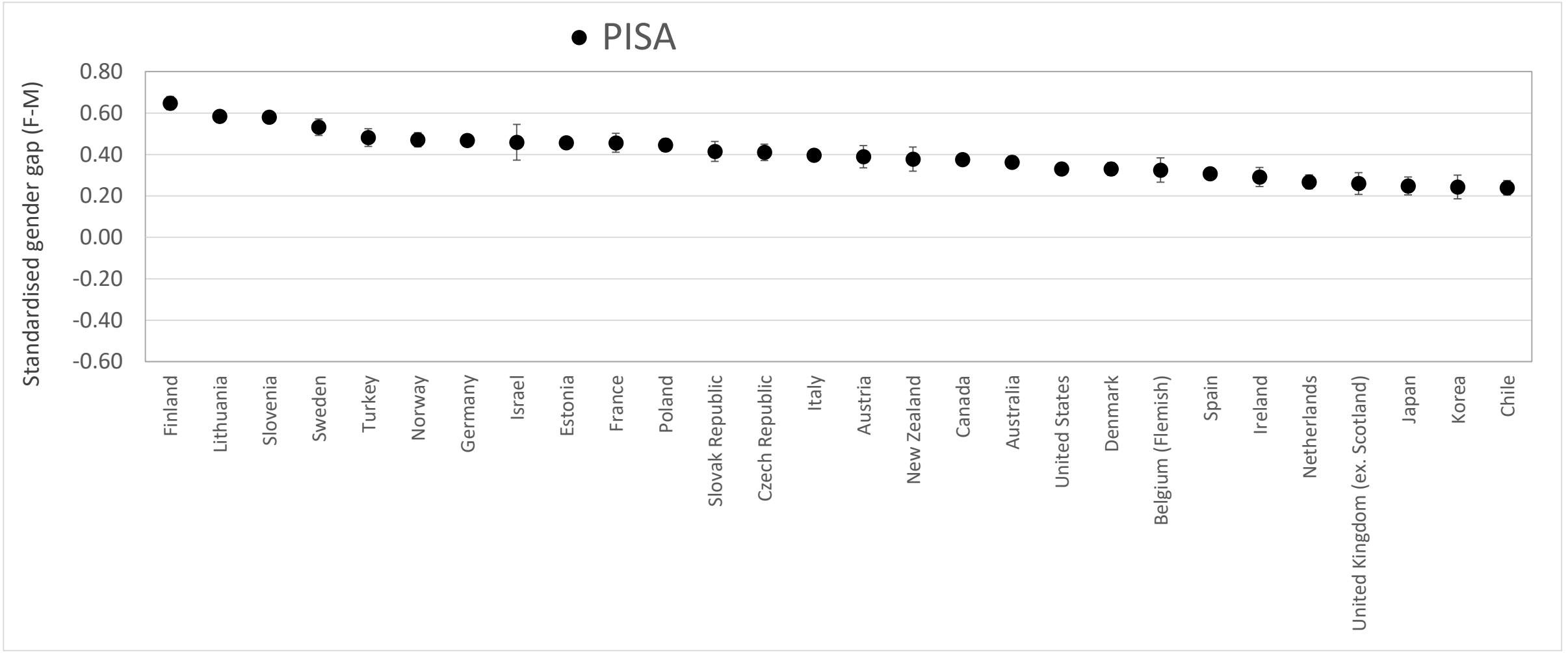
- International large-scale assessments (ILSAs) have been used extensively in academic research and education policy to identify gender gaps in achievement
- Beyond PIAAC: The Programme for International Student Assessment (PISA), the Trends in Mathematics and Science Study (TIMSS), and the Programme for International Reading and Literacy Study (PIRLS)
- Most research examines gender gaps in math
- Increasing research and policy interest in gender gaps in literacy

IOE LONDON

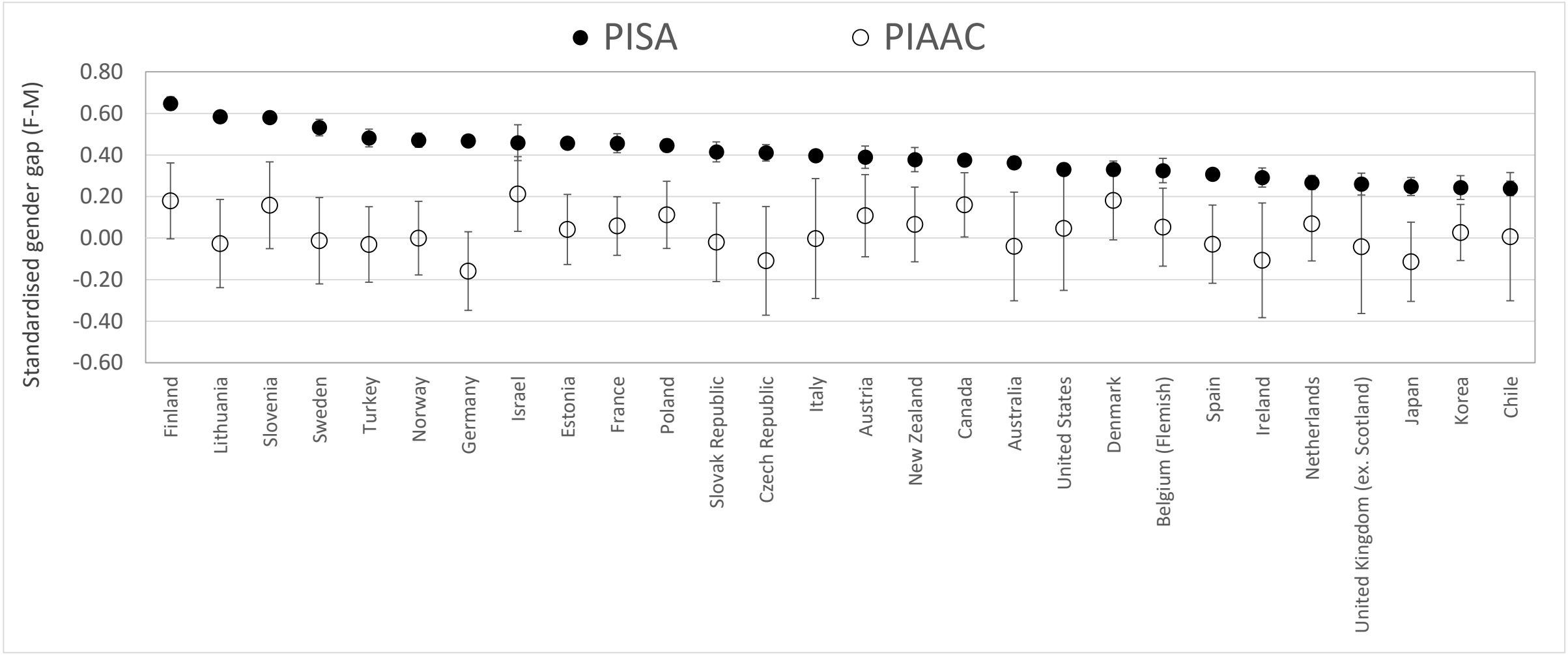
Gender gaps (F-M) according to the PISA 2018 study



Literacy gender gaps around age 15 in PISA 2012



Literacy gender gaps around age 15 in PISA 2012 *and PIAAC*



Why do estimated literacy gender gaps in PISA and PIAAC differ?

- Differences are due to differences in who sits the test
- Differences are due to differences in the tests and gender specific interactions with test material

Test-taking motivation

- PISA and PIAAC are low-stakes. Recent literature questioning between country comparisons
- Expectancy value theory: motivation depends on performance expectancies and task value.
- ‘Motivation’ to take a test – selection effect into PISA and PIAAC (the role of gender specific distributions of abilities)
  - Effect on participation rates
  - Exclusion rates
- Motivation during test-taking –The role of context and text specificities

Differences in test features

- Scoring method (penalties assigned, value of incorrect or missing)
- Test length
- Mode of delivery
- Response format (constructed vs. multiple choice)
- Text structure (continuous – prose – vs. non continuous/mixed texts; narrative content – e.g. cooking vs. football content)

Differences in test features

- Scoring method (penalties assigned, value of incorrect or missing)
- Test length
- Mode of delivery
- Response format (constructed vs choice)
- Text structure (continuou
  narrative content – e.g. c

Until 2012 PISA considered item non-response as wrong while in PIAAC and (PISA 2015 onwards) they are considered as not contributing information to the definition of proficiency. Two sets of estimates are provided: adopting PISA scoring method for both PISA and PIAAC (missing as wrong) and adopting the PIAAC scoring method for both PISA and PIAAC (missing as missing). If males leave more missing (lower motivation) then PISA scoring produces larger gender gaps in PISA in favour of females.T

IOE LONDON

Differences in test features

- Scoring method (penalties assigned, value of incorrect or missing)
- Test length
- Mode o delivery
- Respon onstructed vs. multiple choice)

xts;

PISA is a two-hour **TIMED** assessment composed of 4 clusters of subject specific material designed to take around 30 minutes each to complete. A self-administered questionnaire (30+ minutes) is administered **AFTER** the assessment. PIAAC is **UNTIMED** but designed to take around 40+ minutes to complete (empirically few go above 1 hour and 15 minutes). Two clusters of subject specific material are administered. The questionnaire (30+ minutes) is administered through CAPI **BEFORE** the assessment. We illustrate comparison of performance of males and females at the start of the test (and other elapsed time).

IOE LONDON

Differences in test features

- Scoring method (penalties assigned, value of incorrect or missing)
- Test length
- Mode of delivery
- Response format (construc
- Text structure (continuous – prose
  narrative content – e.g. cooking vs

The main PISA administration was paper-and-pencil in 2012 but in a subset of countries a 40 minutes computer-based assessment was administered. PIAAC was computer-delivered but two thirds of test items were originally developed for paper administration.

IOE LONDON

Differences in test features

- Scoring method (penalties assigned, value of incorrect or missing)
- Test length
- Mode of delivery
- Response format (constructed vs. multiple choice)
- Text structure (continuous – pr
  narrative content – e.g. cooking

Although the literacy frameworks are very similar, in PISA 55% of the test involved constructed responses while in PIAAC there were no constructed responses as such: almost 90% of responses required individuals to click on the correct answer, highlight a piece of text to give an answer, or respond to multiple choice questions. Only in 12% of test questions individuals had to enter text or a number to provide an answer and since answers were computer coded, no extensive writing was involved.

Differences in test features

- Non-continuous texts comprise over 50% of test items in PISA but only 2% of test items in PIAAC. Multiple and mixed texts make up the majority of test items in PIAAC (77% in combined terms) but make up only around a quarter of texts in PISA. Similarly, tasks requiring individuals to access and retrieve information make up only 23% of test items in PISA, but over 55% in PIAAC. Comparisons of mixed texts and Access and retrieve texts is provided.
- Text structure (continuous – prose – vs. non continuous/mixed texts; narrative content – e.g. cooking vs. football content)

Data description

- Comparison between PISA 2012 and PIAAC 2012 (and 2015)
- Focus on PISA 15 and 16-year-olds and PIAAC 16 and 17-year-olds
- Country coverage: 28 countries participating in both studies
- Unbalanced samples: PIAAC was not designed to capture age specific patterns. Effect on SE and CI
- Focus on literacy (very similar framework)
- We report Cohen's d when adjusting for sample restrictions and response rate differences (use of main scales) and average marginal effects (AME) based on % correct when working with item level data
- PIAAC estimates adjusted for differences in item difficulty to account for adaptivity

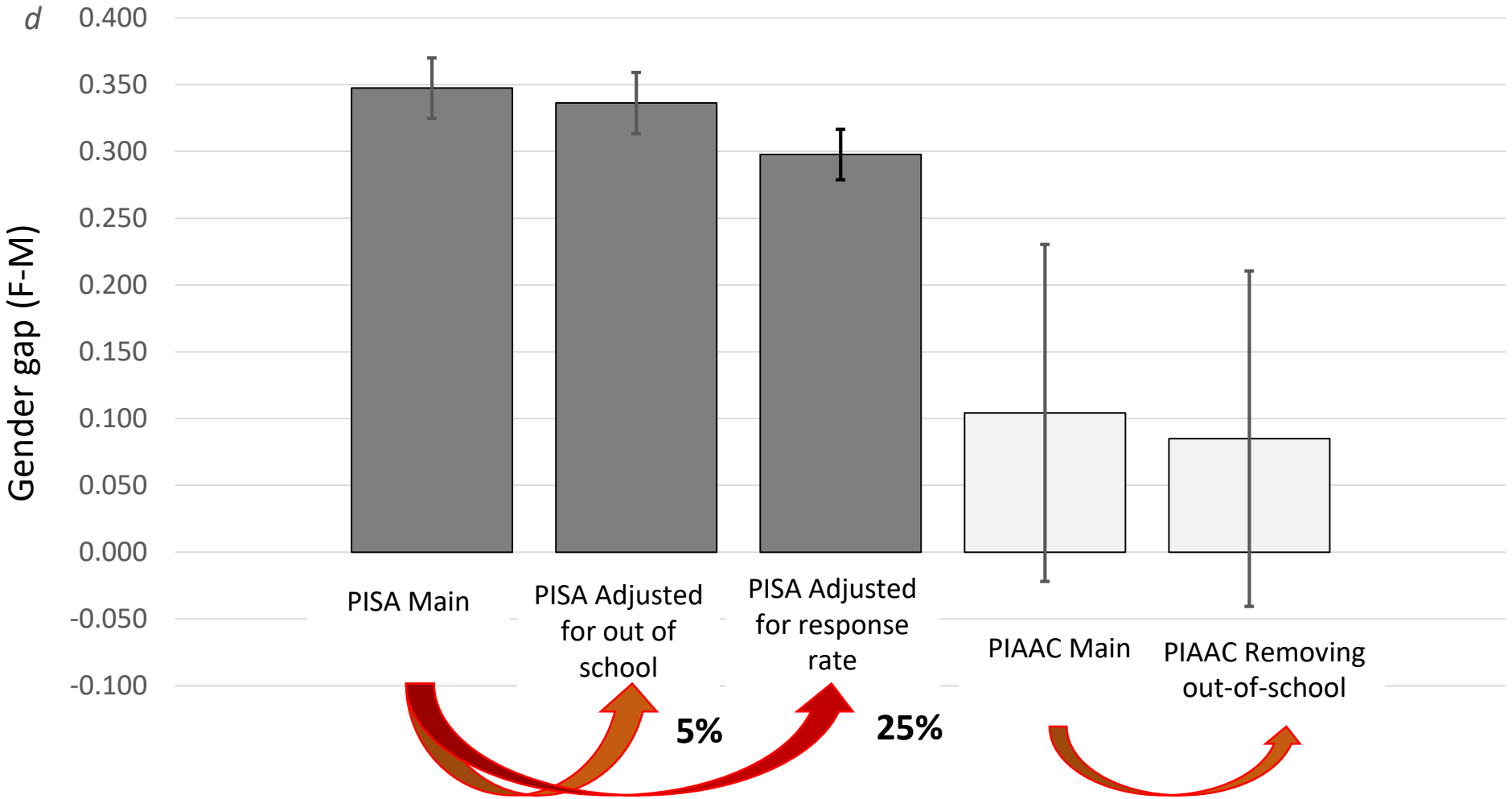
IOE LONDON

Selection effects

- Gender differences at the bottom tail of the distribution are higher
- *The bell and wok phenomenon*
  - Even if gender differences in participation and response rates are the same and males and females are in the same part of the distribution estimates may be affected.
- PISA schooled population. In PIAAC some might have left school. Remove potential drop-outs in PISA (lowest 5% in PISA – worst case assumption).Removal of not in education or training in PIAAC.
- Differences in response rates. (PISA much higher than PIAAC)
- Removal of 25% lowest performers in PISA (worst case scenario)

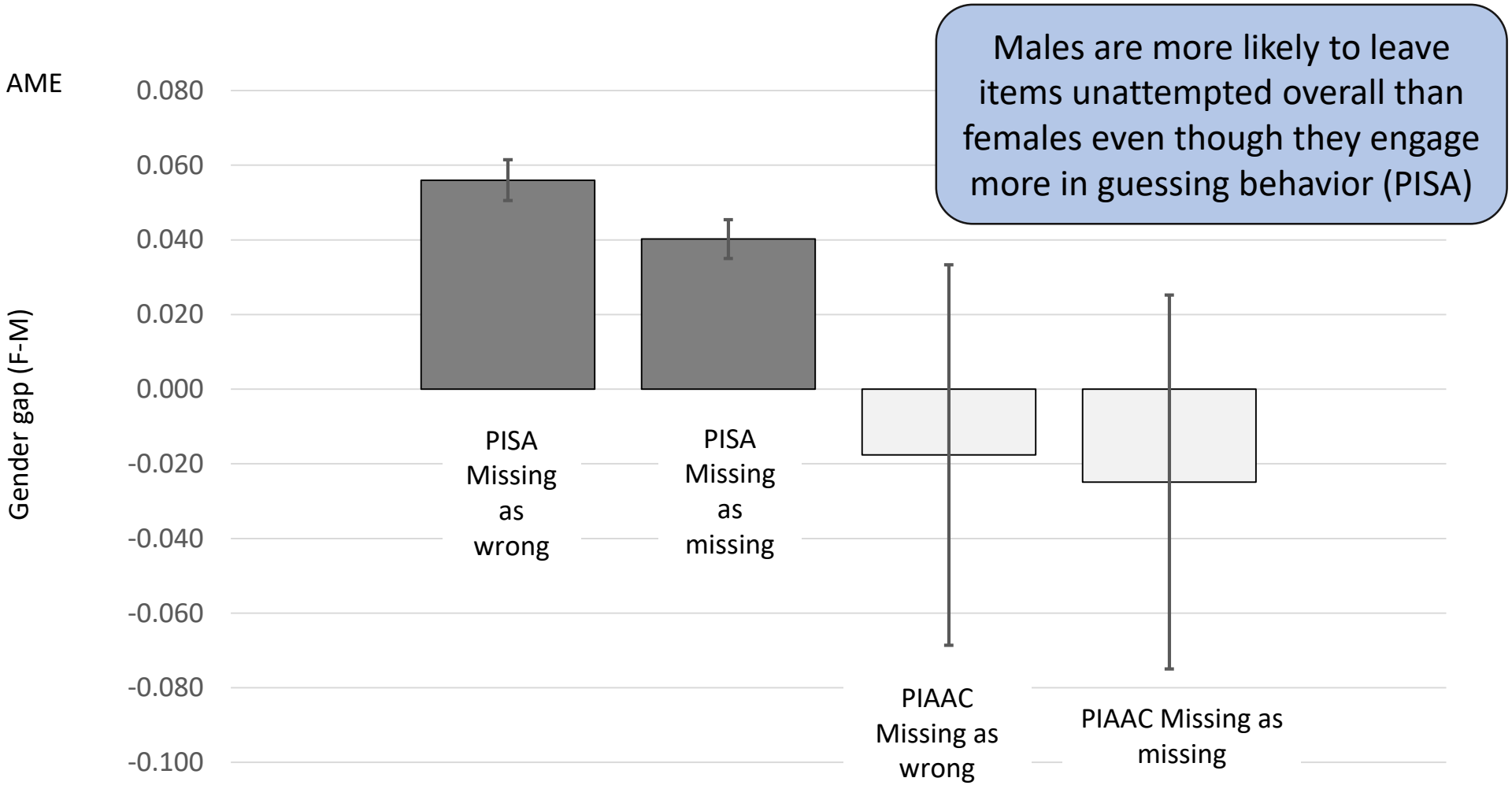
IOE LONDON

Sampling adjustments



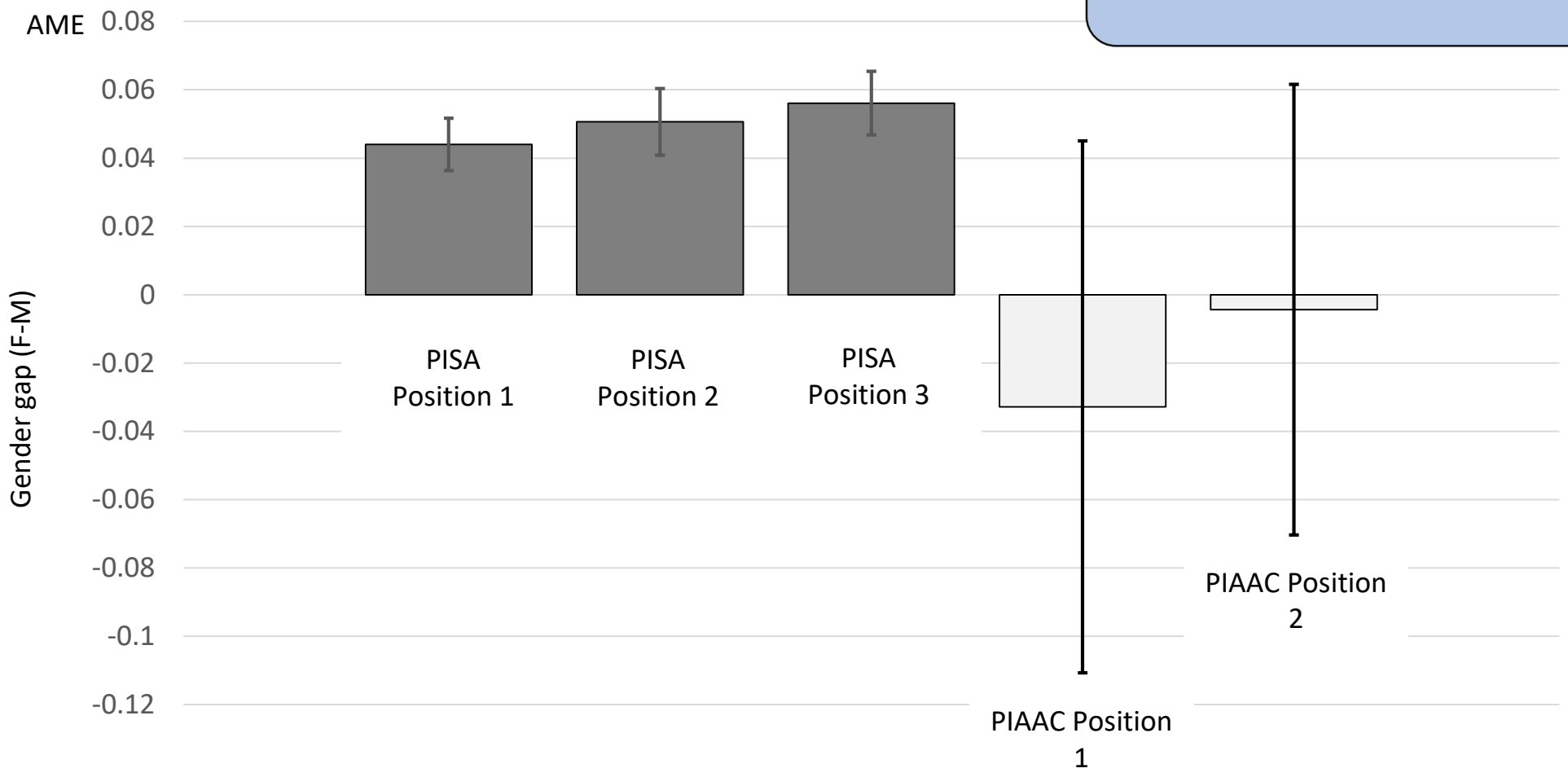
Test construction effects

Scoring method

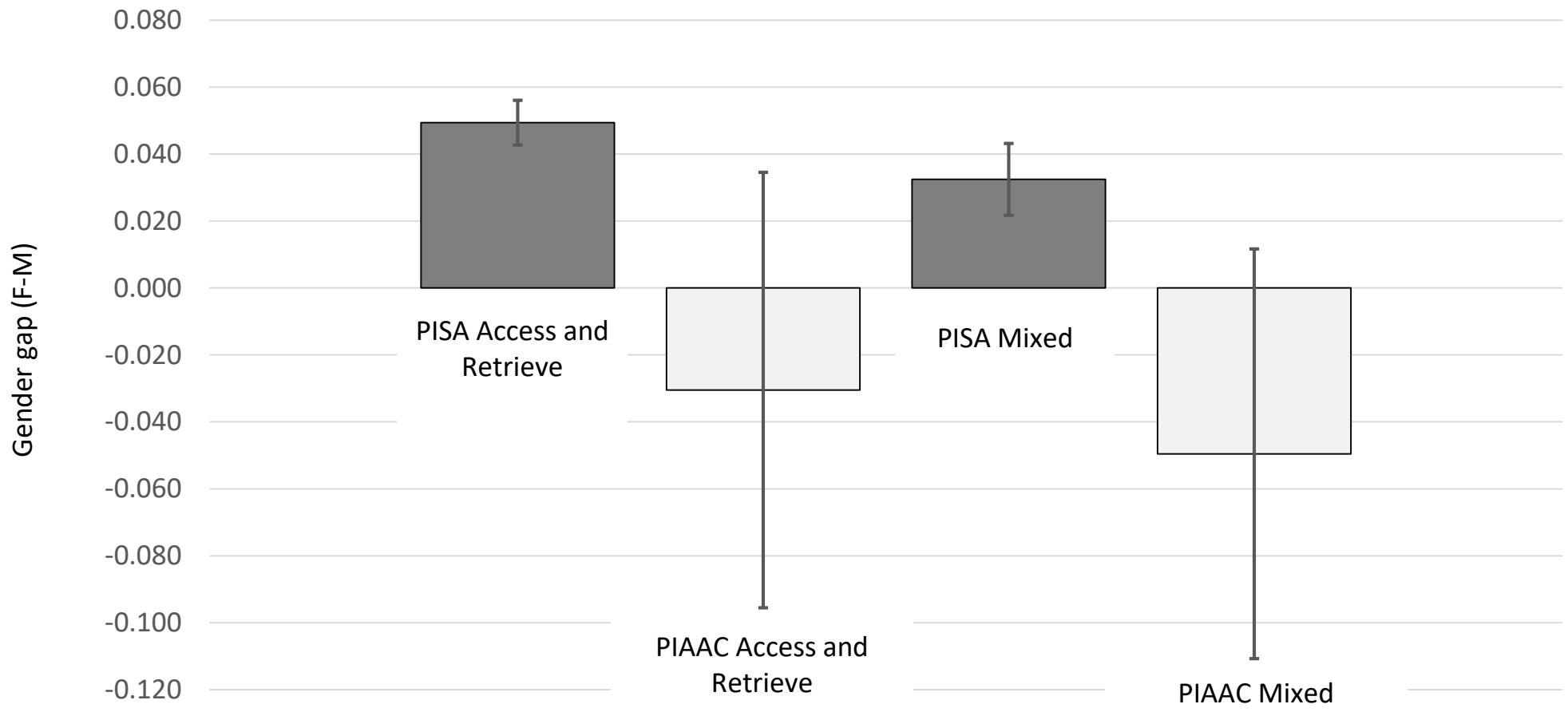


Effect of test length

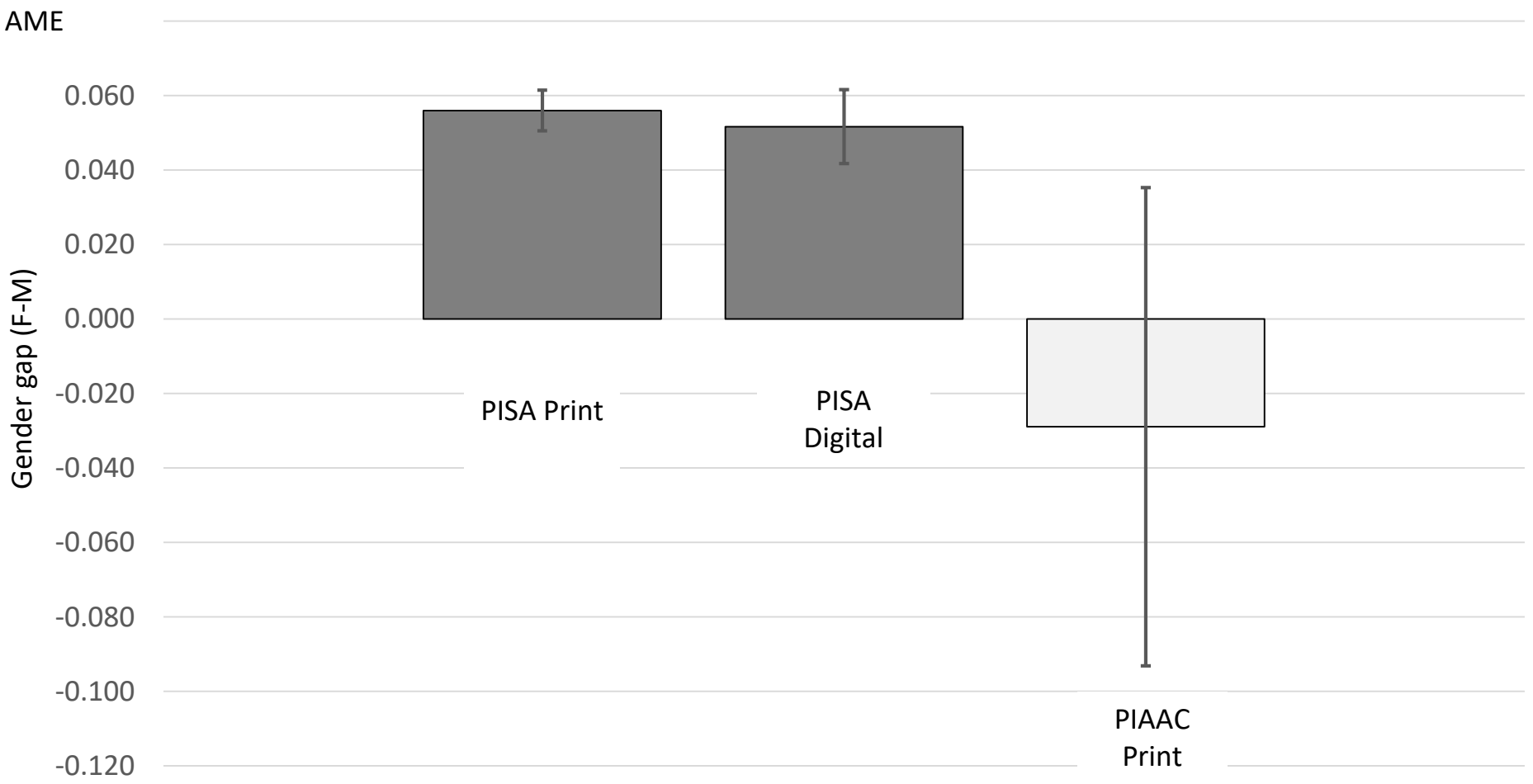
Males are sprinters females are marathon runners



Cognitive processes and text format



Mode of delivery



Full results with formal tests (I)

	PISA		PIAAC		Testing for difference between PISA and PIAAC		
	Standardised gender gap (F-M)	SE	Standardised gender gap (F-M)	SE	Difference (PISA-PIAAC)	SE	t test
PISA main vs. PIAAC adjusted for out of school	0.347	(0.012)	0.085	(0.002)	0.262	(0.012)	22.488
PISA adjusted for out of school vs. PIAAC main	0.336	(0.012)	0.085	(0.002)	0.251	(0.012)	21.201
PISA adjusted for response rate differences vs. PIAAC main	0.298	(0.010)	0.104	(0.002)	0.193	(0.010)	19.390
PISA adjusted for response rate differences vs. PIAAC adjusted for out of school	0.298	(0.010)	0.085	(0.002)	0.213	(0.010)	21.528

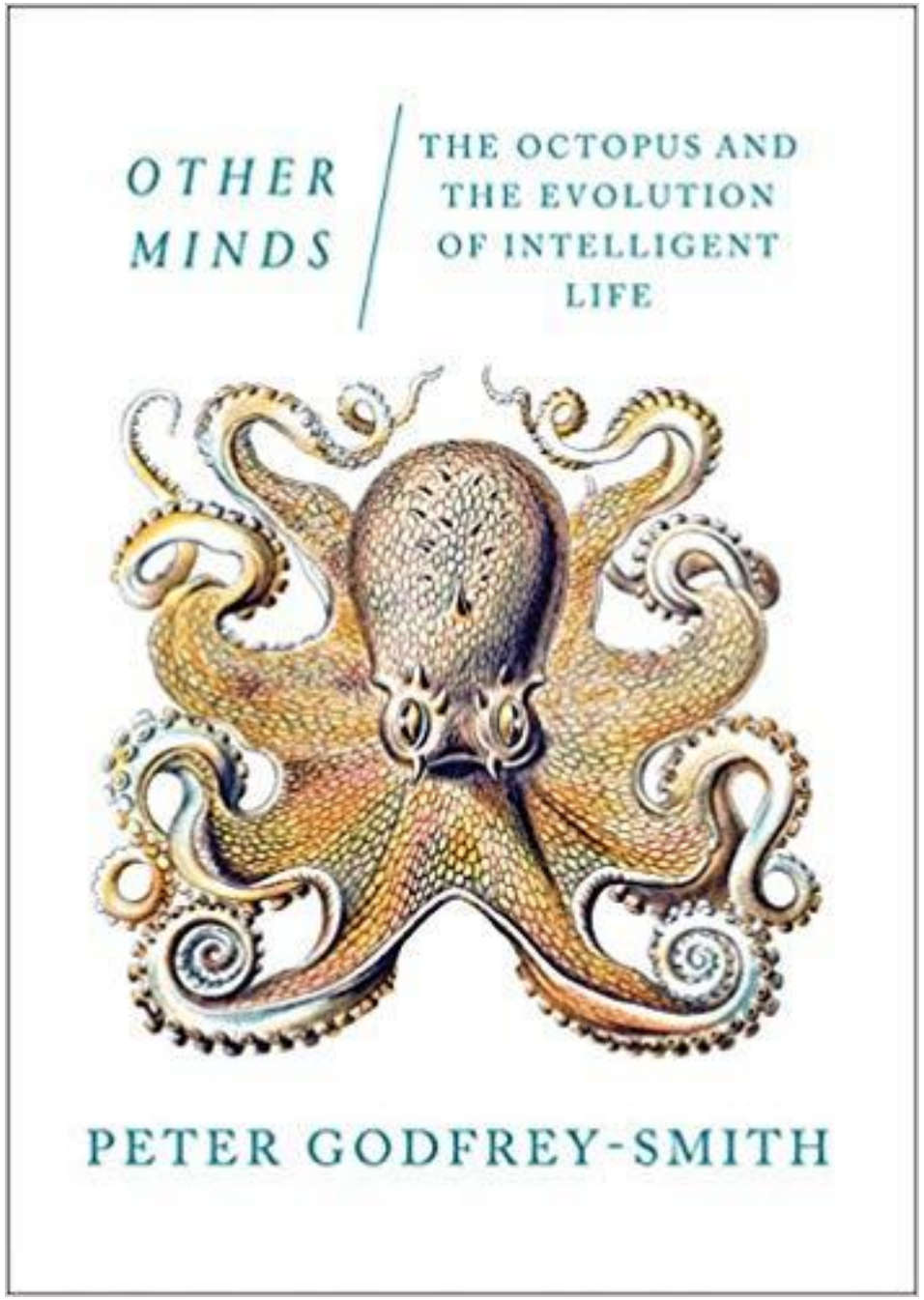
Full results with formal tests (II)

	PISA		PIAAC		Testing for difference between PISA		
	AME (F-M)	SE	AME (F-M)	SE	Difference (PISA-PIAAC)	SE	t test
Scoring effects							
All test: Missing treated as wrong	0.056	0.003	-0.016	0.027	0.072	(0.027)	2.639
All test:Missing treated as missing	0.040	(0.003)	-0.023	0.027	0.063	(0.027)	2.338
Timing effects							
PISA 1 vs. PIAAC 1: missing as wrong	0.044	(0.004)	-0.037	(0.039)	0.081	(0.040)	2.046
PISA 1 vs. PIAAC 1: missing as missing	0.033	(0.004)	-0.048	(0.039)	0.081	(0.039)	2.079
PISA 2 vs. PIAAC 2: missing as wrong	0.051	(0.005)	0.002	(0.036)	0.049	(0.036)	1.358
PISA 2 vs. PIAAC 2: missing as missing	0.035	(0.005)	0.002	(0.036)	0.033	(0.037)	0.904
PISA 1 vs. PIAAC 2: missing as wrong	0.044	(0.004)	0.002	(0.036)	0.042	(0.036)	1.177
PISA 1 vs. PIAAC 2: missing as missing	0.033	(0.004)	0.002	(0.036)	0.031	(0.037)	0.857
Item characteristics effects							
Access and retrieve: missing as wrong	0.049	0.003	-0.030	0.033	0.080	(0.033)	2.394
Access and retrieve: missing as missing	0.034	0.003	-0.040	0.033	0.073	(0.033)	2.235
Mixed: missing as wrong	0.032	0.005	-0.050	0.031	0.082	(0.032)	2.587
Mixed: missing as missing	0.024	0.005	-0.061	0.032	0.085	(0.032)	2.615
Mode of delivery effects							
PISA digital PIAAC all: missing as wrong	0.052	(0.005)	-0.016	0.027	0.067	(0.027)	2.451
PISA digital PIAAC all: missing as missing	0.043	(0.004)	-0.023	(0.027)	0.066	(0.027)	2.436
PISA paper PIAAC paper: missing as wrong	0.056	0.003	-0.040	0.032	0.096	(0.032)	2.953
PISA paper PIAAC paper: missing as missing	0.040	(0.003)	-0.029	0.033	0.069	(0.033)	2.103

Are boys the problem?

…or is the problem that we do not know how to deal with boys?

…or is the problem that boys do not want to have anything to do with us?



Going back to the start…

- Could it be that differences in education and labor markets might have to do with how tests are organized and administered?

What does a school test looks like?

What does a job interview looks like?

Limitations

- It was not possible to identify how much these findings extend to other domains or groups
- Small samples in PIAAC: It was not possible to isolate all effects
- Small samples in PIAAC: It was not possible to isolate for all effects at the same time
- Small samples in PIAAC: It was not possible to study country specific patterns
- Ex post study: items are similar but not the same and are not similar on all characteristics

Take away points

- Proper study: administering PISA in PIAAC conditions and PIAAC in PISA conditions or administering both PISA and PIAAC to the same individuals
- Measurement matters: large-scale assessments should be considered as good measures of well-defined domains, under specific administration conditions. Large, representative samples and comparability across-languages/cultures does not guarantee generalizability of findings to other settings
- Examining why different instruments give different answers to the same question is of substantive interest

IOE LONDON

Thank you!

f.borgonovi@ucl.ac.uk