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
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

Completion rate of upper secondary education by gender (2015)

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

Score point difference

<table>
<thead>
<tr>
<th>Subject</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>30</td>
</tr>
<tr>
<td>Math</td>
<td>-5</td>
</tr>
<tr>
<td>Science</td>
<td>2</td>
</tr>
</tbody>
</table>
Literacy gender gaps around age 15 in PISA 2012

![Graph showing literacy gender gaps around age 15 in PISA 2012]

- Finland
- Lithuania
- Slovenia
- Sweden
- Turkey
- Norway
- Germany
- Israel
- Estonia
- France
- Poland
- Slovak Republic
- Czech Republic
- Italy
- Austria
- New Zealand
- Canada
- Australia
- United States
- Denmark
- Belgium (Flemish)
- Spain
- Ireland
- Netherlands
- United Kingdom (ex. Scotland)
- Japan
- Korea
- Chile

Standardised gender gap (F-M)
Literacy gender gaps around age 15 in PISA 2012 and PIAAC

[Chart showing literacy gender gaps for various countries]
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. multiple choice)
- Text structure (continuous prose vs. non continuous/mixed texts; narrative content – e.g. cooking vs. football content)

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.
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)

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).
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. scientific content)

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.
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)

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
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)
Sampling adjustments

Gender gap (F-M) 

- PISA Main
- PISA Adjusted for out of school
- PISA Adjusted for response rate
- PIAAC Main
- PIAAC Removing out-of-school

-0.100
-0.050
0.000
0.050
0.100
0.150
0.200
0.250
0.300
0.350
0.400

$\Delta$ 5%

$\Delta$ 25%
Test construction effects
Scoring method

Males are more likely to leave items unattempted overall than females even though they engage more in guessing behavior (PISA)
Effect of test length

Males are sprinters females are marathon runners
Cognitive processes and text format
Mode of delivery

The diagram illustrates the gender gap in AME for different modes of delivery:

- **PISA Print**
- **PISA Digital**
- **PIAAC Print**

The y-axis represents the Gender gap (F-M) and the x-axis represents the AME.
## Full results with formal tests (I)

<table>
<thead>
<tr>
<th></th>
<th>PISA main vs. PIAAC adjusted for out of school</th>
<th>PISA adjusted for out of school vs. PIAAC main</th>
<th>PISA adjusted for response rate differences vs. PIAAC main</th>
<th>PISA adjusted for response rate differences vs. PIAAC adjusted for out of school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardised gender gap (F-M)</td>
<td>0.347 (0.012)</td>
<td>0.336 (0.012)</td>
<td>0.298 (0.010)</td>
<td>0.298 (0.010)</td>
</tr>
<tr>
<td>SE</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Standardised gender gap (F-M)</td>
<td>0.085 (0.002)</td>
<td>0.085 (0.002)</td>
<td>0.104 (0.002)</td>
<td>0.085 (0.002)</td>
</tr>
<tr>
<td>SE</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Difference (PISA-PIAAC)</td>
<td>0.262 (0.012)</td>
<td>0.251 (0.012)</td>
<td>0.193 (0.010)</td>
<td>0.213 (0.010)</td>
</tr>
<tr>
<td>SE</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>t test</td>
<td>22.488</td>
<td>21.201</td>
<td>19.390</td>
<td>21.528</td>
</tr>
</tbody>
</table>
Full results with formal tests (II)

<table>
<thead>
<tr>
<th>AME (F-M)</th>
<th>SE</th>
<th>AME (F-M)</th>
<th>SE</th>
<th>SE</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scoring effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All test: Missing treated as wrong</td>
<td>0.056</td>
<td>0.003</td>
<td>-0.016</td>
<td>0.027</td>
<td>0.072</td>
</tr>
<tr>
<td>All test: Missing treated as missing</td>
<td>0.040</td>
<td>(0.003)</td>
<td>-0.023</td>
<td>0.027</td>
<td>0.063</td>
</tr>
<tr>
<td><strong>Timing effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PISA 1 vs. PIAAC 1: missing as wrong</td>
<td>0.044</td>
<td>(0.004)</td>
<td>-0.037</td>
<td>(0.039)</td>
<td>0.081</td>
</tr>
<tr>
<td>PISA 1 vs. PIAAC 1: missing as missing</td>
<td>0.033</td>
<td>(0.004)</td>
<td>-0.048</td>
<td>(0.039)</td>
<td>0.081</td>
</tr>
<tr>
<td>PISA 2 vs. PIAAC 2: missing as wrong</td>
<td>0.051</td>
<td>(0.005)</td>
<td>0.002</td>
<td>(0.036)</td>
<td>0.049</td>
</tr>
<tr>
<td>PISA 2 vs. PIAAC 2: missing as missing</td>
<td>0.035</td>
<td>(0.005)</td>
<td>0.002</td>
<td>(0.036)</td>
<td>0.033</td>
</tr>
<tr>
<td>PISA 1 vs. PIAAC 2: missing as wrong</td>
<td>0.044</td>
<td>(0.004)</td>
<td>0.002</td>
<td>(0.036)</td>
<td>0.042</td>
</tr>
<tr>
<td>PISA 1 vs. PIAAC 2: missing as missing</td>
<td>0.033</td>
<td>(0.004)</td>
<td>0.002</td>
<td>(0.036)</td>
<td>0.031</td>
</tr>
<tr>
<td><strong>Item characteristics effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access and retrieve: missing as wrong</td>
<td>0.049</td>
<td>0.003</td>
<td>-0.030</td>
<td>0.033</td>
<td>0.080</td>
</tr>
<tr>
<td>Access and retrieve: missing as missing</td>
<td>0.034</td>
<td>0.003</td>
<td>-0.040</td>
<td>0.033</td>
<td>0.073</td>
</tr>
<tr>
<td>Mixed: missing as wrong</td>
<td>0.032</td>
<td>0.005</td>
<td>-0.050</td>
<td>0.031</td>
<td>0.082</td>
</tr>
<tr>
<td>Mixed: missing as missing</td>
<td>0.024</td>
<td>0.005</td>
<td>-0.061</td>
<td>0.032</td>
<td>0.085</td>
</tr>
<tr>
<td><strong>Mode of delivery effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PISA digital PIAAC all: missing as wrong</td>
<td>0.052</td>
<td>(0.005)</td>
<td>-0.016</td>
<td>0.027</td>
<td>0.067</td>
</tr>
<tr>
<td>PISA digital PIAAC all: missing as missing</td>
<td>0.043</td>
<td>(0.004)</td>
<td>-0.023</td>
<td>(0.027)</td>
<td>0.066</td>
</tr>
<tr>
<td>PISA paper PIAAC paper: missing as wrong</td>
<td>0.056</td>
<td>0.003</td>
<td>-0.040</td>
<td>0.032</td>
<td>0.096</td>
</tr>
<tr>
<td>PISA paper PIAAC paper: missing as missing</td>
<td>0.040</td>
<td>(0.003)</td>
<td>-0.029</td>
<td>0.033</td>
<td>0.069</td>
</tr>
</tbody>
</table>
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
Thank you!

f.borgonovi@ucl.ac.uk