Christine Sälzer & Fabian Zehner

PISA 2012 in Germany: Exploring Gender Differences in Reading

Conference:
PISA and its meaning for policy and practice around the Baltic Sea, Tallinn, Estonia

Presentation based on:
Gender gap: A classic in Reading Performance

**PISA 2000**

**NAEP 1998 – 2015**

(adapted from NCES, 2015)

(OCED, 2002, S. 147)

(OECD, 2010, S. 59)
Different means = gender gap?

- Comparing means in Reading proficiency disguises useful information
- In fact: big overlap of distributions despite large mean differences (Stanat & Kunter, 2002)
- Mean scores are not representative of their entire distribution

→ Is the „gap“ really a gap?
Gender gap in Germany

- Over the past 6 PISA cycles, the gap in Reading performance between boys and girls was substantial

- 2000: 34 points (OECD, 2002; Stanat & Kunter, 2002)
- 2006: 42 points (OECD, 2007; Drechsel & Artelt, 2007)
- 2009: 40 points (OECD, 2010; Naumann et al., 2010)
- 2015: 21 points (OECD, 2016; Weis, Zehner, Sälzer, Strohmaier, Artelt & Pfost, 2016)
Explanations?

- Our study attempts to explain what may be behind this constantly found gap
- Can students‘ open-ended responses to PISA Reading items be used to explain the gender gap?
- Some earlier studies suggest that girls‘ and boys‘ reading proficiency could be an effect of item format x motivation; but effect sizes are very small (Schwabe, McElvany & Trendtel, 2015)
- Differential reading motivation may be due to differential interests (OECD, 2015):
  - Girls prefer fiction (+ 19%) and magazines (+ 14%)
  - Boys prefer newspapers (+ 7%) and comics (+ 10%)
- Strong predictors of reading literacy: Reading engagement and strategies
Our study

• Taking into account that the score distributions of girls and boys in Reading overlap
• A closer look into the responses to open-ended questions:
  Do girls’ and boys’ responses differ? Can we identify response types?
  Types are then described as „girls‘ type“ and „boys‘ type“ → prototypes
• Using automatic coding (natural language processing, NLP) for analyzing the students‘ open-ended responses
  Identifying similarities and differences
• Assumptions about cognitive processes behind a response
Theory: Cognitive model

- Propositional information from the stimulus text…

  …is processed to situation model (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983) → holistic model describing which cognitive processes are involved while readers try to understand a text

  …distinguishing micro- and macropropositions:
  Micropropositions carry information on text entities like persons or their relations
  Macropropositions contain higher-order information such as a paragraph‘s gist

- Analyzing the students‘ responses, we determine which and how many propositions they selected and how response features can be mapped to cognitive processes (Graesser & Franklin, 1990; Graesser & Murachver, 1985; Graesser & Clark, 1985)
Automatic system used

- Approach recently published (Zehner, Sälzer & Goldhammer, 2016)
Automatic system used

Example: Starting with a short text response ...

[a] [girl] [fall] [into] [and] [wander] [through] [a] [fantasy] [world] /

... to a numerical representation of its semantics ... (LSA; Deerwester et al., 1990)

... up to the automatic code
Material and data

- PISA 2012 sample of Germany (continued with the PISA 2015 data)
- 8 Items assessing Reading literacy
- PISA coding guides to decide which parts of a response are relevant or irrelevant

<table>
<thead>
<tr>
<th>Item</th>
<th>Category</th>
<th>Aspect</th>
<th>Correct</th>
<th>n (%f)</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Explain Protagonist’s Feeling</td>
<td>WHY&lt;s&gt;</td>
<td>B</td>
<td>83%</td>
<td>4,152</td>
<td>12.3</td>
</tr>
<tr>
<td>2. Evaluate Statement</td>
<td>ENABLE&lt;a&gt;</td>
<td>C</td>
<td>43%</td>
<td>4,234</td>
<td>15.6</td>
</tr>
<tr>
<td>3. Interpret the Author’s Intention</td>
<td>SIG&lt;a&gt;</td>
<td>B</td>
<td>10%</td>
<td>4,234</td>
<td>12.5</td>
</tr>
<tr>
<td>4. List Recall</td>
<td>CON&lt;s&gt;</td>
<td>A</td>
<td>59%</td>
<td>4,786</td>
<td>13.6</td>
</tr>
<tr>
<td>5. Evaluate Stylistic Element</td>
<td>HOW&lt;s&gt;</td>
<td>C</td>
<td>56%</td>
<td>4,012</td>
<td>10.7</td>
</tr>
<tr>
<td>6. Verbal Production</td>
<td>WHN&lt;a&gt;</td>
<td>B</td>
<td>80%</td>
<td>4,404</td>
<td>11.4</td>
</tr>
<tr>
<td>7. Select and Judge</td>
<td>ENABLE&lt;s&gt;</td>
<td>C</td>
<td>68%</td>
<td>4,136</td>
<td>10.6</td>
</tr>
<tr>
<td>8. Explain Story Element</td>
<td>CON&lt;e&gt;</td>
<td>B</td>
<td>69%</td>
<td>4,229</td>
<td>12.5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>59%</td>
<td>33,604</td>
<td>12.6</td>
</tr>
</tbody>
</table>

- A = Access & Retrieve, B = Integrate & Interpret, C = Reflect & Evaluate, M = Uncertainty & Data, S = Explain Phenomena Scientifically according to PISA framework (OECD, 2013)
- f = percentage of girls
- c = average word count for non-empty responses (SD)
Analyses (condensed)

1. Proposition Entity Count (PEC): Number of PEs

   - How many micro- and macropropositions has the student used in the response?

   - How many of those propositions were relevant, how many were irrelevant for a correct response?
Analyses (condensed)

1. Gender type: Responses were semantically grouped (clusters) and distinguished in responses of the „girl type“ and responses of the „boy type“
   → Gender type instead of gender split

   → Response is taken as an indicator of a response type („girl“, „boy“) independent from the students‘ gender

   → number of clusters determined by human-computer agreement

   → for each cluster, ratio boys : girls determined plus 95%-CI (Oranje, 2006; Wilson, 1927)

   → types with a significantly dominating gender included
Analyses: Summary

• Contrasting **two extreme groups** of student responses

• Distinguishing types according to the **gender majority** in each cluster (i.e., a cluster dominated by girls will be a „girl-type“ response cluster)

• Student responses are used as a database for drawing **conclusions about cognitive processes** lying behind these responses \(\rightarrow\) principle of IRT and many large-scale student assessments
### Results (selection) I

**Proposition Entity Count for Gender Type and Correctness**

<table>
<thead>
<tr>
<th>Item</th>
<th>$\beta_{gt}$</th>
<th>$\beta_c$</th>
<th>$\beta_{gt*c}$</th>
<th>df1, df2</th>
<th>$F$ (df1, df2)</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Explain Protagonist’s Feeling</td>
<td><strong>-0.84</strong></td>
<td><strong>4.9</strong></td>
<td><strong>0.46</strong></td>
<td>3, 1065</td>
<td><strong>227.50</strong></td>
<td><strong>.389</strong></td>
</tr>
<tr>
<td>2. Evaluate Statement</td>
<td>-0.37</td>
<td>0.19</td>
<td>-0.27</td>
<td>3, 378</td>
<td><strong>97.18</strong></td>
<td><strong>.431</strong></td>
</tr>
<tr>
<td>3. Interpret the Author’s Intention</td>
<td>-0.42</td>
<td>-0.04</td>
<td>0.26</td>
<td>3, 668</td>
<td><strong>35.76</strong></td>
<td><strong>.135</strong></td>
</tr>
<tr>
<td>4. List Recall</td>
<td>-0.51</td>
<td>-0.14</td>
<td>NA$^a$</td>
<td>2, 2600</td>
<td><strong>207.20</strong></td>
<td><strong>.137</strong></td>
</tr>
<tr>
<td>5. Evaluate Stylistic Element</td>
<td>-0.33</td>
<td>0.35</td>
<td>0.01</td>
<td>3, 735</td>
<td><strong>89.36</strong></td>
<td><strong>.264</strong></td>
</tr>
<tr>
<td>6. Verbal Production</td>
<td><strong>-0.31</strong></td>
<td><strong>2.9</strong></td>
<td><strong>-0.57</strong></td>
<td>3, 802</td>
<td><strong>163.60</strong></td>
<td><strong>.377</strong></td>
</tr>
<tr>
<td>7. Select and Judge</td>
<td>-0.37</td>
<td>0.24</td>
<td>-0.11</td>
<td>3, 1018</td>
<td><strong>124.10</strong></td>
<td><strong>.266</strong></td>
</tr>
<tr>
<td>8. Explain Story Element</td>
<td>-0.38</td>
<td>0.10</td>
<td>-0.01</td>
<td>3, 952</td>
<td><strong>69.55</strong></td>
<td><strong>.177</strong></td>
</tr>
</tbody>
</table>

*bold = significant ($\alpha = .05$), $gt =$ gender type (1 = girl type, 2 = boy type), $c =$ correct response

$^a$ no girl-specific responses

- Responses of the **boy-type** contain **less PEs** than responses of the **girl-type**
Results (selection) II

Proposition Entity Count for Gender Type and Correctness

Gender Type and Correctness Main Effects

- 4-List Recall
  - Girl-Specific Types
  - Boy-Specific Types

- 5-Evaluate Stylistic Element
  - Girl-Specific Types
  - Boy-Specific Types

- 7-Select and Judge
  - Girl-Specific Types
  - Boy-Specific Types
Results (selection) III

Micro and Relevance Measures
Results (selection) IV

Micro and Relevance Measures
Results: Summary

Different cognitive types dominated by genders instead of homogeneously separated genders

Boy Type
- parsimonious selection of PEs (controlled for correctness)
- if response is correct, efficacy is worthwhile, but:
- type is relatively consistent across correct and incorrect responses \(\rightarrow\) less stable situation model
- flawed reconstruction (Kintsch & van Dijk, 1978)

Girl Type
- Elements in the situation model are juggled with quite high flexibility (PEC & relevance)
- micro level where required
- macro level where required
- Question focus and category are identified more easily

Cognitive types could be an effect of different reading strategies (Artelt, Naumann & Schneider, 2010)
Limitations

- **Data limited to Germany** only at this point
  → Joint research project with appr. 12 other PISA NCs using further languages

- **Relevance measures** are based on coding guides which are not empirically exhaustive (Zehner, Goldhammer & Sälzer, 2015)
  → relevance is partly underestimated

- **Operationalizations** yield only rough implementation of the situation model
  → improvements through progress in NLP
  → further features are possible

- Expand theory and operationalizations to allow **text responses** to be a new source of information for further content-related research questions
Thank you for your kind attention!

Contact details:
Dr. Christine Sälzer
christine.saelzer@tum.de
References


