PIAAC Innovations: Facilitating deeper understanding of survey data

The use of process data, bioinformatics, and modern psychometrics

Matthias von Davier, ETS
Characteristics of Surveys of Student and Adult Populations

• International large scale surveys of skills are low-stakes assessments
• We cannot ‘make’ test takers perform their best, we can only ask, motivate, and provide incentives for participation
• Any test data suffers from a mix of test taker motivational states, some are motivated, are interested, or do like the assessment, so do so to a lower extent
Some Examples of How Low Effort may Look Like

In PIAAC:

Question 1. Circle the telephone number given in the advertisement below.

SGIH
SUPPORT GROUP FOR THE INTEGRATION OF THE HOMELESS
Av. Duque de Loule, nº 44, 1º Floor - 1050 Lisboa
Tel. (01)3338200

NOT in PIAAC:

3. Find x.

Here it is
Our Approach is Grounded in Behavioral Research

- Theory of planned behavior (e.g., Fishbein & Ajzen, 1975)
- Achievement Motivation - Motivation to succeed (e.g., Eccles, Wigfield & Schiefele, 1998)
  - Ability related beliefs, Expectancy of success, Usefulness, Importance, Interest
- Effort moderated test taking (e.g., Wise & DeMars, 2005)
An Effort to Take Test-Taking-Effort in Account

• If effort is an issue, are the results of PIAAC (and PISA, and TIMSS and NAEP...) still useful?
• Effort is an issue in all testing, but cannot be fully controlled
• Even in college admission tests there will be test taker who do not ‘give it all’
• PIAAC reduced the issue using incentives, which helps (e.g., Braun, Kirsch & Yamamoto, 2011)
• Our research aims to further reduce impact and improve reliability, validity and comparability
3 Examples for Methodological Innovations in PIAAC

• Timing data and effort moderation models (a variation on Wise & DeMars, 2005) – with Jon Weeks and Kentaro Yamamoto

• Bioinformatics and careless response deviations in computerized text-selection items – with Jana Sukkarieh and Kentaro Yamamoto

• Extreme response styles, test taking effort, and associations with proficiency measures – with Lale Khorramdel
Timing data and effort moderation

- Response time (speed) and accuracy have been investigated since computers were used for testing. Formal models that separate are beginning to see more and more application.
- Wise & DeMars (2005) suggest using response time to determine whether responses are valid.
  - They suggest eliminating all responses produced in less than X seconds.
  - This produces missing data where there were codes for incorrect / correct for rapid answers.
Timing data and effort moderation

- Omitted responses are item level non-response due to choice of respondents to not provide an answer. Omitted responses are frequently found in low stakes assessments, more so for constructed response items than for multiple choice items -> Effort may play a role here
- Treating omitted responses as ignorable missing data is one option
- Assuming that all omitted responses indicate lack of skill is another option
- In PIAAC, we have response times to examine these choices
Timing data and effort moderation
Timing data and effort moderation

- Wise & DeMars (2005) treat rapid (correct/incorrect/omitted) responses as ignorable missing data.
- Traditional testing programs treat missing data as wrong responses. This was shown to introduce bias (O’Muircheartaigh & Moustaki, 1999; Moustaki & Knott, 2000; Glas & Pimentel, 2008; Holman & Glas, 2005; Rose, von Davier & Xu, 2010).
- In this project, we follow an approach that avoids extremes. Rapid non-response (omission or skipping of an item) are treated as missing responses, while for ‘slow’ missing responses we assume that test-takers engaged in solving the item and were unable to provide a response.
Timing data and effort moderation

Table 1: Percentage of rapid omissions & not reached vs. regular non-response. Literacy and Numeracy are combined for the paper based assessment, but kept separate for the Computer based assessment (CBA) due to the larger number of items.

<table>
<thead>
<tr>
<th></th>
<th>Paper based L&amp;N</th>
<th>CBA-Literacy</th>
<th>CBA- Numeracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Omitted -&gt; Wrong</td>
<td>7</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Direct right/wrong</td>
<td>88</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>All scored responses</td>
<td>95</td>
<td>98</td>
<td>98</td>
</tr>
</tbody>
</table>
Summary: Timing data and effort moderation

- The assignment of wrong responses to omissions over 5 seconds in the computer based path produces similar levels of recoding as in the paper based assessment.

- Overall, the level of omissions (recoded or not) is lower in the computer based assessment compared to the paper based assessment.

- Future studies (focusing on next PIAAC, PISA etc. assessments) will aim at using model based treatments of speed/response propensity and incorporate more item and respondent variables.
Bioinformatics and Response Deviations

Jana Sukkarieh, Matthias von Davier, & Kentaro Yamamoto, ETS
A number of computer-based literacy items used the response mode of highlighting in PIAAC.

The scoring task was to determine whether respondents highlight a section of the stimulus text that was asked for in the questions accompanying the stimulus material.

Example: “Please indicate by highlighting an example of a response mode used in PIAAC”

The multilingual nature of PIAAC made it impossible to utilize traditional tools from natural language processing, as most of these tools are language dependent.
A number of computer-based literacy items used the response mode of highlighting in PIAAC.

The scoring task was to determine whether respondents highlight a section of the stimulus text that was asked for in the questions accompanying the stimulus material.

Example: “Please indicate by highlighting an example of a response mode used in PIAAC”

The multilingual nature of PIAAC made it impossible to utilize traditional tools from natural language processing, as most of these tools are language dependent.
Bioinformatics and Careless Response Deviations

- A number of computer-based literacy items used the response mode of highlighting in PIAAC.
- The scoring task was to determine whether respondents highlight a section of the stimulus text that was asked for in the questions accompanying the stimulus material.
- Example: “Please indicate by highlighting an example of a response mode used in PIAAC.”
- The multilingual nature of PIAAC made it impossible to utilize traditional tools from natural language processing, as most of these tools are language dependent.
Bioinformatics and Careless Response Deviations

- A number of computer-based literacy items used the response mode of highlighting in PIAAC.
- The scoring task was to determine whether respondents highlight a section of the stimulus text that was asked for in the questions accompanying the stimulus material.
- Example: “Please indicate by highlighting an example of a response mode used in PIAAC.”
- The multilingual nature of PIAAC made it impossible to utilize traditional tools from natural language processing, as most of these tools are language dependent.
Bioinformatics and Careless Response Deviations


Among the 33 ‘highlighting’ items that were used in PIAAC, 16 items were found to provide improved scores using a ‘forgiveness’ rule that allowed some careless omission or insertion.

For 17 out of 33 items, the scoring rules as used in the computer-based platform were sufficient.

The ‘forgiveness’ rules (omission/insertion) threshold allowed on average for a buffer of about 8% of the expected length of the correct response.
## Bioinformatics and Careless Response Deviations

Overview of the scoring adjustments to allow for some level of omission or insertion in the 33 highlighting items used in the CBA branch of PIAAC:

<table>
<thead>
<tr>
<th>Threshold</th>
<th>0%</th>
<th>3%</th>
<th>5%</th>
<th>7.5%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. of items</td>
<td>17</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Cumulative</td>
<td>17</td>
<td>25</td>
<td>27</td>
<td>28</td>
<td>33</td>
</tr>
</tbody>
</table>
Summary: Bioinformatics and Careless Response Deviations

- Whether a response deviation is due to a lack of effort or a potential lack of fine motor coordination, a too rigorous scoring rule applied to highlighting responses appears to be undesirable.

- Some level of response deviation is easily detected as irrelevant by human scorers, while automated scoring falls short unless some algorithmic adjustment is made that utilizes pattern recognition or artificial intelligence.

- Gene matching and protein sequencing algorithms such as the LCS are well-developed tools that can be utilized for this analytic task in large scale technology based skill surveys.
Extreme Responses and Test Taking Effort
Lale Khorramdel & Matthias von Davier, ETS
**PIAAC background items assessing numeracy skill use in the workplace**

<table>
<thead>
<tr>
<th>Stem</th>
<th>In your [job or last job], how often [do or did] you usually...</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>calculate prices, costs or budgets?</td>
</tr>
<tr>
<td>(b)</td>
<td>use or calculate fractions, decimals or percentages?</td>
</tr>
<tr>
<td>(c)</td>
<td>use a calculator – either hand-held or computer based?</td>
</tr>
<tr>
<td>(d)</td>
<td>you usually prepare charts, graphs or tables?</td>
</tr>
<tr>
<td>(e)</td>
<td>use simple algebra or formulas?</td>
</tr>
<tr>
<td>(f)</td>
<td>use more advanced math or statistics such as ...</td>
</tr>
</tbody>
</table>
### Extreme Responses and Test Taking Effort

Example of responses of four respondents and their total score, using codes (never=0), (less than monthly=1), (less than once per week=2), (few days per week=3), (every day=4):

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Item (a)</th>
<th>Item (b)</th>
<th>Item (c)</th>
<th>Item (d)</th>
<th>Item (e)</th>
<th>Item (f)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Arielle</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Mickey</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Mulan</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>
Extreme (and other) response styles have been discussed for a long time in the literature on causes of nuisance factors in self reports. Approaches to identify and correct for these effects include:

- Statistical Modeling (*Mixture IRT*, Multidimensional IRT Models, Multinomial Tree Models)
- Changes in Response Formats (e.g. use of forced choice items)
- Use of Additional Variables (e.g. anchoring vignettes)
Extreme Responses and Test Taking Effort

Mixture distribution models were first used by Newcomb (1886) and Pearson (1894) in analyses of biological data, and have been used in Psychometrics for 25+ years (Yamamoto, 1987; Mislevy & Verhelst, 1990; Kelderman & Macready, 1990; Rost & von Davier, 1992, …)

- Mixture models help identify homogeneous populations
- Have been used for identification of motivation effects, response styles, test speededness, strategy differences, and faking, and typological versus factorial individual differences
Mixtures of Response Types in Skill Use Data
Mixtures of Response Types in Skill Use Data

... I have the same plot for some 8 or so more countries ...

- There are 3 populations/classes emerging in these analyses
  - ‘right’ class that allows quantitative skill use differences to be identified between test takers
  - ‘left’ class that contains respondents for whom the questions don’t apply
  - ‘extreme’ class that frequently selects 0 or 4 (extreme negative and positive) responses
Mixtures of Response Types in Skill Use Data
Mixtures of Response Types in Skill Use Data

- The model with 3 populations/classes fits the data better than models without this distinction.
- This implies that the data on skill use gives us two levels of information:
  - Response type: Extreme, Not applicable, Quantifiable
  - Skill use intensity and variety
- Similar results were found on questionnaires based on the Five Factor Model.
Validity Evidence on Response Types in Skill Use Data

Expected PIAAC numeracy scale scores in the 3 subpopulations across 8 countries

<table>
<thead>
<tr>
<th>Numeracy</th>
<th>CNT8</th>
<th>CNT7</th>
<th>CNT6</th>
<th>CNT5</th>
<th>CNT4</th>
<th>CNT3</th>
<th>CNT2</th>
<th>CNT1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch</td>
<td>280.45</td>
<td>278.65</td>
<td>271.72</td>
<td>295.73</td>
<td>292.74</td>
<td>283.52</td>
<td>300.47</td>
<td>274.33</td>
</tr>
<tr>
<td>Resp.Set</td>
<td>256.43</td>
<td>268.15</td>
<td>254.03</td>
<td>292.19</td>
<td>288.20</td>
<td>269.68</td>
<td>293.56</td>
<td>264.64</td>
</tr>
<tr>
<td>Zero</td>
<td>234.18</td>
<td>240.03</td>
<td>232.50</td>
<td>255.85</td>
<td>268.00</td>
<td>246.53</td>
<td>265.86</td>
<td>246.82</td>
</tr>
</tbody>
</table>
## Validity Evidence on Response Types in Skill Use Data

Expected PIAAC literacy scale scores in the 3 subpopulations across 8 countries

<table>
<thead>
<tr>
<th>Literacy</th>
<th>CNT8</th>
<th>CNT7</th>
<th>CNT6</th>
<th>CNT5</th>
<th>CNT4</th>
<th>CNT3</th>
<th>CNT2</th>
<th>CNT1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch</td>
<td>291.29</td>
<td>273.62</td>
<td>274.80</td>
<td>291.71</td>
<td>285.55</td>
<td>289.72</td>
<td>298.01</td>
<td>282.86</td>
</tr>
<tr>
<td>Resp.Set</td>
<td>274.21</td>
<td>265.29</td>
<td>256.75</td>
<td>283.58</td>
<td>282.38</td>
<td>276.41</td>
<td>290.01</td>
<td>273.82</td>
</tr>
<tr>
<td>Zero</td>
<td>254.23</td>
<td>245.77</td>
<td>240.51</td>
<td>255.66</td>
<td>266.20</td>
<td>260.26</td>
<td>269.51</td>
<td>256.72</td>
</tr>
</tbody>
</table>
Regression of Skill use on Numeracy given Response Style

![Graph showing the regression of skill use on numeracy given response style.](image)
Regression of Skill use on Numeracy given Response Style

- Skill use response style is a significant predictor of performance on literacy and numeracy.
- When using country specific regression models we find:
  1. Response types: Extreme & Not applicable associated with lower average literacy / numeracy.
  2. Skill-use intensity and variety is also a significant predictor.
- Note that this approach is based on extracting 2 indicators from a single (short) skill-use scale.
Outlook: Response effort, styles, and assessment

- Test taking effort affects every assessment result
- Modern psychometrics and computer based assessments provide multiple sources of relevant data
- More research on improving indicators of engagement, test-taking effort and persistence is needed
- This will result in a better understanding of differences between groups of respondents
- PIAAC does, and ICT-based PISA and NAEP data will provide rich data for this type of research