Maximizing Validity of Self-report Measures for socio-emotional outcomes through rigorous survey design and psychometric analyses
Outline

• What self-report data are used for
• Reference groups and other challenges
• A case for more rigorous survey design principles
  • Question design, selection, interpretation
Good questionnaire design may be an art, but it’s definitely a science.

P.S. Thinking that good art does not require skill and discipline is an insult to the artist.
What self-reports are used for

Demographic Information
- What is your gender?
- How old are you?
- What is your race/ethnicity?
- What's your income?

Attitudes and Beliefs

Other factual/objective information

J. Bertling, 10/14/2022
What self-reports are used for

- Demographic Information
  - Did you have breakfast today?
  - Last week, how many days did you exercise?
  - Last year, how many times did you exercise in a typical week?
  - This school year, how often did your math teacher ask you to solve mathematics problems without computing anything?

- Attitudes and Beliefs
- Other factual/objective information

J. Bertling, 10/14/2022
What self-reports are used for

How much do you agree or disagree with the statement “I finish whatever I begin.”?

On a scale from 1-10, how satisfied are you with your life these days?

How much do you think people change whether they are good in math?
What the general public sees

Well-being at school and at home

23% of students reported being victims of an act of bullying at least a few times a month

Less than 15% of students in Korea, the Netherlands, Portugal and Chinese Taipei reported this

8 in 10 students expressed anti-bullying attitudes, such as

- It is a wrong thing to join in bullying
- It is a good thing to help a student who can defend themselves

Bullying measure is based on student self-report questions.

Less positive views of math

More positive views of math

Each bubble represents a group of students

Student views measure is based on self-report questions.

Percentage of students reporting attitudes in these ranges

<table>
<thead>
<tr>
<th></th>
<th>LOW</th>
<th>MODERATE</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>11%</td>
<td>39%</td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>

Average score and percentage of eighth-grade students assessed in NAEP mathematics, by students’ views of mathematics index scores: 2015

Strong math performance

Poor math performance

https://www.nationsreportcard.gov/sq_students_views_2015/
What the general public sees (Cont’d)

Strength of the socio-economic gradient and reading performance

Equity measure is based on self-report questions.

Each diamond is a country

Strong reading performance

Poor reading performance

Low degree of equity

High degree of equity

https://www.oecd-ilibrary.org/education/2018_pisa_key_findings_4ed4125e.pdf?expires=1637346268&id=id&accname=guest&checksum=51CE9BA08B4859B8080716B2142F554C
Self-report Surveys and Social and Emotional Skills

How is the person thinking about themselves? What attitudes does the person have?

What behaviors is the person engaging in? What skills is the person demonstrating?

Demographic Information

Attitudes and Beliefs

Other factual/objective information

J. Bertling, 10/14/2022
Social and Emotional Skills

- Perseverance
- Self Control
- Stress Resistance
- Emotional Control
- Assertiveness
- Curiosity
- Cooperation
- Trust
- Empathy
Reference groups and other challenges

- The indicator works differently across individuals from different groups
- The indicator is not valid
- The indicator is not reliable
ARE YOU SURE THE DATA YOU GAVE ME IS CORRECT?

I’VE BEEN GIVING YOU INCORRECT DATA FOR YEARS. THIS IS THE FIRST TIME YOU’VE ASKED.

WHAT?

I SAID THE DATA IS TOTALLY ACCURATE.
It all starts with the item – There are many ways to end up with bad data

- Can respondents accurately calibrate their answer?
- Do respondents have necessary level of self-awareness?
- Can respondents remember?
- Are respondents willing to disclose accurate information?
- Do all respondents understand the question in the same way?

Demographics  | Factual Information/behavioral reports  | Attitudes/Beliefs

(J. Bertling, 10/14/2022)
Respondent Behaviors that can cause bias

- **Acquiescence** – General tendency to agree with statements
- **Extreme Response** – General tendency to pick extreme response options
- **Midpoint Response** – General tendency to choose the middle
- **Patterns of disengaged responding**, e.g. straightlining
- **Reference group** – Tendency to calibrate one’s answer relative to a reference group
### Acquiescence

**ST034**

**Thinking about your school: to what extent do you agree with the following statements?**

*(Please select one response in each row.)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel like an outsider (or left out of things) at school.</td>
<td>☐ 01</td>
<td>☒ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>I make friends easily at school.</td>
<td>☒ 01</td>
<td>☐ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>I feel like I belong at school.</td>
<td>☒ 01</td>
<td>☐ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>I feel awkward and out of place in my school.</td>
<td>☐ 01</td>
<td>☒ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>Other students seem to like me.</td>
<td>☒ 01</td>
<td>☐ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>I feel lonely at school.</td>
<td>☐ 01</td>
<td>☒ 02</td>
<td>☐ 03</td>
<td>☐ 04</td>
</tr>
</tbody>
</table>
**Extreme Responding**

**Thinking about your school: to what extent do you agree with the following statements?**

*(Please select one response in each row.)*

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel like an outsider (or left out of things) at school.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>I make friends easily at school.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>I feel like I belong at school.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>I feel awkward and out of place in my school.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>Other students seem to like me.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>I feel lonely at school.</td>
<td>01</td>
<td>02</td>
<td>03</td>
<td>04</td>
</tr>
</tbody>
</table>

J. Bertling, 10/14/2022
Reference Group Bias

“Big Fish Little Pond Effect”

<table>
<thead>
<tr>
<th>Reference Group Statements</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST034Q05TA Other students seem to like me.</td>
<td>☐ 01</td>
<td>☒ 02</td>
<td>☒ 03</td>
<td>☐ 04</td>
</tr>
<tr>
<td>ST034Q06TA I feel lonely at school.</td>
<td>☒ 01</td>
<td>☐ 02</td>
<td>☒ 03</td>
<td>☐ 04</td>
</tr>
</tbody>
</table>

J. Bertling, 10/14/2022
Differential Item Functioning

• An item shows DIF is the probability of a certain response is (partially) dependent on a person variable that is not theoretically related to the construct that is being measured.

• DIF is an unexpected difference in item difficulty between groups due to something other than the construct of interest

• DIF is a systematic effect, not just additional random (measurement) error

• DIF has impact on validity: A test score is not messaging the same thing across groups
Thoughtful principled survey design can minimize impact of these issues.
Survey Questionnaires Development Phases in Large-Scale Assessments
Basics of Measurement Theory

• “Classical Test Theory” describes the effects of measurement error on test scores
  • Error = not a mistake, but inconsistencies caused by random influences on test scores

• Item Response Theory models the probability of a correct response to an item (or agreement with a statement), conditional on the level of the construct measured (latent trait, theta)
$X = T + E$

- “observed” scores $X$: values assigned on the basis of measurement instrument used
- True score $T$: hypothetical entity the respondent would obtain if measurements were free of all error
- Error $E$: assumed to be random
Item Characteristic Curves

- How many response categories do your questions have?
- Can you reasonably make the assumption that every item in the test is equally indicative of the latent trait you are measuring?
- How likely is guessing to be a problem on your test?
Reliability and Validity in Everyday language

• **Reliability:** “You should get a similar score if you repeat the measurement”
• How accurately are we measuring WHATEVER we measure

• **Validity:** “A test is valid if it measures what it is supposed to measure”
• Do we measure the right thing?
Reliable Not Valid
Low Validity Low Reliability
Not Reliable Not Valid
Both Reliable and Valid
“The Standards”

- “Validity refers to the **degree to which evidence and theory support the interpretations** of test scores for proposed uses of tests.”
- “The process of validation involves **accumulating relevant evidence** to provide a sound scientific basis for the proposed score interpretations.”
- “It is the **interpretations of test scores** for proposed uses that are evaluated, not the test itself.”
- “Statements about validity should refer to particular interpretations **for specified uses**. It is incorrect to use the unqualified phrase ‘the validity of the test.’”
Validity or Reliability?

- Zumbo and Chan (2014): psychological scientists tend to report relatively little validity evidence and focus much more on other psychometric properties, most importantly reliability.
- Simplest explanation: providing reliability evidence is relatively easy, whereas providing validity evidence is very hard.
Common methods to assess Reliability

- **Internal consistency** (Cronbach’s Alpha): average correlation between all items of the test
- **Split-half** (coefficient of stability): correlation between two “halves” of the test
- **Test-retest** (coefficient of stability): correlation between scores on the same test collected at two different times
- **Standard Error of Measurement**: Standard deviation of an individual’s observed scores around their true score; derived as total score standard deviation * Square root of (1 minus reliability coefficient)
- **Information functions**: In IRT reliability is estimated with regard to the latent trait, not the observed test score. IRT allows for the estimation of different reliabilities for different test scores
Item Information Function

- Test information function indicates the precision of the theta estimates
- Test information = sum of item information functions
- Standard error of the estimate = inverse of the square root of the test information function:

\[ \frac{1}{\sqrt{I(\theta)}}. \]
Kinds of Validity Evidence

“the Standards”
1. Test content
2. Response processes
3. Internal structure
4. Relationships with other variables
5. Consequences of testing
Big Five Personality Assessment

- Based on BFI-44

<table>
<thead>
<tr>
<th>Big Five Dimensions</th>
<th>Facet (and correlated trait adjective)</th>
</tr>
</thead>
</table>
| Extraversion vs. introversion | Gregariousness (sociable)  
Assertiveness (forceful)  
Activity (energetic)  
Excitement-seeking (adventurous)  
Warmth (outgoing) |
| Agreeableness vs. antagonism | Trust (forgiving)  
Straightforwardness (not demanding)  
Altruism (warm)  
Compliance (not stubborn)  
Modesty (not show-off)  
Tender-mindedness (sympathetic) |
| Conscientiousness vs. lack of direction | Competence (efficient)  
Order (organized)  
Dutifulness (not careless)  
Achievement striving (thorough)  
Self-discipline (not lazy)  
Deliberation (not impulsive) |
| Neuroticism vs. emotional stability | Anxiety (tense)  
Angry hostility (irritable)  
Depression (not contented)  
Self-consciousness (shy)  
Impulsiveness (moody)  
Vulnerability (not self-confident) |
| Openness vs. closedness to experience | Ideas (curious)  
Fantasy (imaginative)  
Aesthetics (artistic)  
Actions (wide interests)  
Feelings (excitable)  
Values (unconventional) |

https://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Personality-BigFiveInventory.pdf
Nomological Nets
How many factors?

- Some key questions to guide you in interpreting a scree plot
  - Is there a clear “elbow”?
  - How many components have eigenvalues “above the elbow”?
  - How many components have eigenvalues above 1?
What are the factors?
Interpreting factor loadings

- Does the item have a loading on any factor that is $>|.2|$?
- Does the item have loadings $>|.2|$ on more than one factor?
- Are items for the same theoretical constructs loading on the same factor?
- Is there a substantive possible explanation for cross-loadings?
- Is there a substantive possible explanation for additional factors (beyond those expected based on theory)?
It all starts with the individual item.
"Duckworth, Peterson, Matthews, and Kelly (2007) introduced the construct of grit, defined as trait-level perseverance and passion for long-term goals." (p. 166)
# Conscientiousness via BFI-10

How well do the following statements describe your personality?
I see myself as someone who ...

<table>
<thead>
<tr>
<th>Statement</th>
<th>Disagree strongly</th>
<th>Disagree a little</th>
<th>Neither agree nor disagree</th>
<th>Agree a little</th>
<th>Agree strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>... is reserved</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... is generally trusting</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... tends to be lazy</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... is relaxed, handles stress well</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... has few artistic interests</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... is outgoing, sociable</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... tends to find fault with others</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... does a thorough job</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... gets nervous easily</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... has an active imagination</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

---

## Brief Report

**Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German**

**Beatrice Rammstedt**, Oliver P. John

Abstract

To provide a measure of the Big Five for contexts in which participant time is severely limited, we abbreviated the Big Five Inventory (BFI-44) to a 10-item version, the BFI-10. To permit its use in cross-cultural research, the BFI-10 was developed simultaneously in several samples in both English and German. Results focus on the psychometric characteristics of the 2-item scales on the BFI-10, including their part-whole correlations with the BFI-44 scales, retest reliability, structural validity, convergent validity with the NEO-PI-R and its facets, and external validity using peer ratings. Overall, results indicate that the BFI-10 scales retain significant levels of reliability and validity. Thus, reducing the items of the BFI-44 to less than a fourth yielded effect sizes that were lower than those for the full BFI-44 but still sufficient for research settings with truly limited time constraints.

© 2006 Elsevier Inc. All rights reserved.

**Keywords**: Big Five personality dimensions, Five-Factor Model, Short measures, Reliability, Validity, Test construction
Conscientiousness via BFI-44

Conscientiousness vs. lack of direction

<table>
<thead>
<tr>
<th>Conscientiousness</th>
<th>Lack of direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence (efficient)</td>
<td>Order (organized)</td>
</tr>
<tr>
<td>Order (organized)</td>
<td>Dutifulness (not careless)</td>
</tr>
<tr>
<td>Dutifulness (not careless)</td>
<td>Achievement striving (thorough)</td>
</tr>
<tr>
<td>Achievement striving (thorough)</td>
<td>Self-discipline (not lazy)</td>
</tr>
<tr>
<td>Self-discipline (not lazy)</td>
<td>Deliberation (not impulsive)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>20. Has an active imagination</td>
<td>42. Likes to cooperate with others</td>
</tr>
<tr>
<td>21. Tends to be quiet</td>
<td>43. Is easily distracted</td>
</tr>
<tr>
<td>22. Is generally trusting</td>
<td></td>
</tr>
<tr>
<td>23. Tends to be lazy</td>
<td></td>
</tr>
<tr>
<td>24. Is emotionally stable, not easily upset</td>
<td></td>
</tr>
<tr>
<td>25. Is inventive</td>
<td></td>
</tr>
<tr>
<td>26. Has an assertive personality</td>
<td></td>
</tr>
<tr>
<td>27. Can be cold and aloof</td>
<td></td>
</tr>
<tr>
<td>28. Perseveres until the task is finished</td>
<td></td>
</tr>
<tr>
<td>29. Can be moody</td>
<td></td>
</tr>
<tr>
<td>30. Values artistic, aesthetic experiences</td>
<td></td>
</tr>
<tr>
<td>31. Is sometimes shy, inhibited</td>
<td></td>
</tr>
<tr>
<td>32. Is considerate and kind to almost everyone</td>
<td></td>
</tr>
<tr>
<td>33. Does things efficiently</td>
<td></td>
</tr>
<tr>
<td>34. Remains calm in tense situations</td>
<td></td>
</tr>
<tr>
<td>35. Prefers work that is routine</td>
<td></td>
</tr>
<tr>
<td>36. Is outgoing, sociable</td>
<td></td>
</tr>
<tr>
<td>37. Is sometimes rude to others</td>
<td></td>
</tr>
<tr>
<td>38. Makes plans and follows through with them</td>
<td></td>
</tr>
<tr>
<td>39. Gets nervous easily</td>
<td></td>
</tr>
<tr>
<td>40. Likes to reflect, play with ideas</td>
<td></td>
</tr>
<tr>
<td>41. Has few artistic interests</td>
<td></td>
</tr>
</tbody>
</table>

Scoring:

BFI scale scoring ("R" denotes reverse-scored items):

Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

J. Bertling, 10/14/2022
1. **How many questions** will you need for one topic?
2. How can you choose the most appropriate **response options**?
3. Which item **format works** best?
How many questions will you need for one topic?
How many questions you will need depends on the type of topic you are targeting.

**Observable**
- How many computers with internet access are available in school X?

**Not directly observable**
- How perseverant is student X?
- What is student X’s socio-economic status?

**Can be measured with one question**

Need multiple indicators to measure the construct

Statistical aggregation in to an “index” for reporting

For example:
- Student finishes tasks she starts.
- Student does not give up after making mistakes.
- Student applies more effort when tasks become difficult.
- …
If you aim to measure a construct, develop multiple questions.

- **Too few questions** lead to:
  - Low reliability
  - Poor construct representation

- **Build in a “buffer”** for question selection after pre-testing and piloting
  - *Rule of thumb:* Start with twice the number of questions that you would like to report on later
  - Develop 10 questions if you aim to measure a construct with a 5-item index

<table>
<thead>
<tr>
<th></th>
<th># Items</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BELONG</td>
<td>3</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.687</td>
</tr>
<tr>
<td>COGACT</td>
<td>3</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.705</td>
</tr>
<tr>
<td>MATHWKETH</td>
<td>3</td>
<td>0.682</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.744</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.785</td>
</tr>
<tr>
<td>MATHEFF</td>
<td>3</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.742</td>
</tr>
<tr>
<td>MATHBEH</td>
<td>3</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.540</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.593</td>
</tr>
<tr>
<td>FAMCON</td>
<td>3</td>
<td>0.476</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.549</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.605</td>
</tr>
</tbody>
</table>

From: Bertling & Weeks (2020)
How can you choose the most appropriate response options?
Don’t default to “agree-disagree” options.

The most commonly used response options in surveys are: Strongly disagree / Disagree / (Neither agree nor disagree) / Agree / Strongly agree.

What do you want to know?

- Yes or no?
- How often?
- How many?
- How much?

- Absolute frequency
- Relative frequency
- Exact count
- Approximate
- X do you have in Y?
- …do you like X?
- …do you agree with X?

- Invite social desirable responding
- Response styles
- Not every question is about agreement!

Don’t make the “Task” too hard for the respondent!
Response options are essential part of the question, don’t make them an afterthought

• Think about which response options will provide you with the best data for your question of interest.

• Consider:
  • Ease of responding
  • Utility for reporting

Don’t use too many or too few response options

Avoid vague terms like “sometimes”, “often”, “rarely”

Use labels for all scale points

Cover the entire range of possible answers

Offer an “out” if question may be not applicable to certain respondents
Which item format works best?
Most commonly used item formats in survey questionnaires are “Discrete” and “Matrix” questions.

- We tested the impact of the item format in a NAEP special study.

Do you think you would be able to write sentences and paragraphs using a computer?

- I definitely can’t.
- I probably can’t.
- I probably can.
- I definitely can.

Do you think you would be able to edit text using a computer?

- I definitely can’t.
- I probably can’t.
- I probably can.
- I definitely can.

---

<table>
<thead>
<tr>
<th></th>
<th>I definitely can’t.</th>
<th>I probably can’t.</th>
<th>I probably can.</th>
<th>I definitely can.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Write sentences and paragraphs using a computer</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>b. Edit text using a computer</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>c. Use a touchscreen on a computer, tablet, or smartphone</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>d. Look up the meaning of a word using a computer</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>e. Draw a picture using a computer</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Same questions in two different formats
Findings from NAEP study with 4th graders:

- No notable differences in statistical data.
- Matrix questions take less time to answer.
- Don’t combine questions that cannot be combined.
- Make sure the stem matches the items.
- Don’t make matrices too long.

But:

Does it matter where contextual cues are placed in survey questions?

When you study geography, how much does each of the following statements describe a person like you?

<table>
<thead>
<tr>
<th></th>
<th>Version A</th>
<th>Version B</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>I want to learn as much as possible about geography in my class.</td>
<td>I want to learn as much as possible in my class.</td>
</tr>
<tr>
<td>b.</td>
<td>I want to master a lot of new geography skills in my class.</td>
<td>I want to master a lot of new skills in my class.</td>
</tr>
<tr>
<td>c.</td>
<td>I want to become a better geography student this year.</td>
<td>I want to become a better student this year.</td>
</tr>
<tr>
<td>d.</td>
<td>I want to understand as much as I can about geography in my class.</td>
<td>I want to understand as much as I can in my class.</td>
</tr>
</tbody>
</table>
Does it matter where contextual cues are placed in survey questions?

Qureshi, Gill, Alegre, & Bertling (2018)

<table>
<thead>
<tr>
<th>Average difference</th>
<th>Performance Goals</th>
<th>Civics</th>
<th>Not at all like me</th>
<th>Exactly like me</th>
<th>Geography</th>
<th>Not at all like me</th>
<th>Exactly like me</th>
<th>U.S. History</th>
<th>Not at all like me</th>
<th>Exactly like me</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-3.03%</td>
<td>6.92%</td>
<td>-3.68%</td>
<td>6.11%</td>
<td>-2.12%</td>
<td>4.41%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-7.58%</td>
<td>22.63%</td>
<td>-7.18%</td>
<td>19.43%</td>
<td>-4.80%</td>
<td>13.56%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Positive values indicate more frequent endorsement of response option in stem-context version. Negative values indicate more frequent endorsement of response option in item-context version.

When the cue is in the stem substantially more students choose “Exactly like me” and fewer students choose “Not at all like me” than when each item is contextualized.
### Within-construct Matrix Sampling

#### Traditional Design: Each student answers all questions for each construct

<table>
<thead>
<tr>
<th>Construct</th>
<th>Question</th>
<th>Not at all confident</th>
<th>Not very confident</th>
<th>Confident</th>
<th>Very confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>How confident do you feel about having to do the following mathematics tasks? (Please select one response in each row.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Working out from a train timetable how long it would take to get from one place to another</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating how much more expensive a computer would be after adding tax</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating how many square metres of tiles you need to cover a floor</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Understanding scientific tables presented in an article</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $6x^2 + 3 = 29$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Finding the actual distance between two places on a map with a 1:10,000 scale</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $2(x^2) = (x + 3)(x - 3)$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating the power consumption of an electronic appliance per week</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $2x + 5 = 17$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

#### Innovative Design: Each student answers a subset of questions for each construct

<table>
<thead>
<tr>
<th>Construct</th>
<th>Question</th>
<th>Not at all confident</th>
<th>Not very confident</th>
<th>Confident</th>
<th>Very confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>How confident do you feel about having to do the following mathematics tasks? (Please select one response in each row.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Working out from a train timetable how long it would take to get from one place to another</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating how much more expensive a computer would be after adding tax</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating how many square metres of tiles you need to cover a floor</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Understanding scientific tables presented in an article</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $6x^2 + 3 = 29$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Finding the actual distance between two places on a map with a 1:10,000 scale</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $2(x^2) = (x + 3)(x - 3)$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Calculating the power consumption of an electronic appliance per week</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>Solving an equation like $2x + 5 = 17$</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

J. Bertling, 10/14/2022
Why >five items per construct?

Matrix Sampling:
Random selection of 5 items per student

Fixed Scale Shortening:
Administration of “5 best” items to every student

PISA 2012
Math Self-Efficacy

PISA 2015
Sense of Belonging

Bertling & Weeks, 2018; Bertling et al., 2020
Feasibility confirmed in PISA 2022 FT
Example: Self-efficacy

<table>
<thead>
<tr>
<th>Item</th>
<th>Item_ID</th>
<th>Ncat</th>
<th>Slope</th>
<th>Difficulty</th>
<th>Step1</th>
<th>Step2</th>
<th>Step3</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>ST290Q01WA</td>
<td>4</td>
<td>0.93</td>
<td>0.05</td>
<td>1.08</td>
<td>0.08</td>
<td>-1.15</td>
</tr>
<tr>
<td>02</td>
<td>ST290Q02WA</td>
<td>4</td>
<td>0.90</td>
<td>0.13</td>
<td>1.14</td>
<td>-0.01</td>
<td>-1.13</td>
</tr>
<tr>
<td>03</td>
<td>ST290Q03WA</td>
<td>4</td>
<td>1.16</td>
<td>0.09</td>
<td>1.09</td>
<td>0.00</td>
<td>-1.09</td>
</tr>
<tr>
<td>04</td>
<td>ST290Q04WA</td>
<td>4</td>
<td>0.85</td>
<td>0.28</td>
<td>1.34</td>
<td>0.01</td>
<td>-1.35</td>
</tr>
<tr>
<td>05</td>
<td>ST290Q05WA</td>
<td>4</td>
<td>1.30</td>
<td>-0.21</td>
<td>0.85</td>
<td>0.02</td>
<td>-0.67</td>
</tr>
<tr>
<td>06</td>
<td>ST290Q06WA</td>
<td>4</td>
<td>0.79</td>
<td>0.34</td>
<td>1.24</td>
<td>-0.11</td>
<td>-1.13</td>
</tr>
<tr>
<td>07</td>
<td>ST290Q07WA</td>
<td>4</td>
<td>1.14</td>
<td>-0.21</td>
<td>0.85</td>
<td>0.03</td>
<td>-0.89</td>
</tr>
<tr>
<td>08</td>
<td>ST290Q08WA</td>
<td>4</td>
<td>0.68</td>
<td>0.38</td>
<td>1.30</td>
<td>-0.06</td>
<td>-1.24</td>
</tr>
<tr>
<td>09</td>
<td>ST290Q09WA</td>
<td>4</td>
<td>1.26</td>
<td>-0.41</td>
<td>0.69</td>
<td>0.06</td>
<td>-0.74</td>
</tr>
</tbody>
</table>

Based on PISA 2022 Field Trial Data

J. Bertling, 10/14/2022
Summing it up
There are many ways to end up with bad data

- Can respondents accurately calibrate their answer?
- Do respondents have necessary level of self-awareness?
- Can respondents remember?
- Are respondents willing to disclose accurate information?
- Do all respondents understand the question in the same way?
Question Understanding

• Plain language
• Avoid multi-barreled questions
• Clear response options
• If several languages are used, conduct a translatability review
• Test in cognitive interviews
Willingness to disclose

- Keep burden low
- Confidentiality
- Low-stakes
- Avoid asking about sensitive topics
- Test in cognitive interviews
Ability to Recall

- Simple wording
- Don’t ask anything you wouldn’t remember yourself
- Avoid too vague response options
- Test in cognitive interviews
How to judge the quality of YOUR questionnaire

Conceptually:
- Do the items “sound” like what you’re interested in measuring? (content validity)
- Example for a problem: You’re interested in how student’s time management has changed after an intervention. You’re using a short conscientiousness questionnaire because you know that time management is one facet of conscientiousness. However, the two items used in your conscientiousness scale is based on are “I finish whatever I begin” and “People see me as a trustworthy person”.
- Can you clearly interpret data from your questionnaire (e.g. double-barreledness?, response options)
- Have you minimized potential for biases by principled item design?

Empirically:
- Item response frequency patterns – is there variation?
- Reliability – does your scale have acceptable level of reliability, e.g. >.80
- Validity – do scores on your scale correlate with scores from other questionnaires claiming to measure the same?
- Scalability and DIF
How to build reliable scales

• Include enough items
• Include enough good items
• Include enough good items with sufficiently different surface characteristics
Item Writing Checklist

- Do you want to measure a construct (i.e., something that cannot be directly observed)?
  - If yes, have you developed a sufficient number of items to begin with? (10 is usually a good starting point)

- Have you thought about which response options (ROs) maximize ease for respondent and utility for reporting? (Think of alternatives to agreement! Remember that the response options you choose will determine what you can possibly report later on.)
  - Do you have a good reason to use fewer or more than 5 ROs? (If not, 5 is usually a good number.)
  - Do your response options cover the entire range? (If not, add ROs.)
  - Are all items applicable to all respondents? (If not, add ROs.)
  - Do all your response options have verbal labels? (If not, add labels.)
  - Are you using vague labels that could be replaced with more specific ones?

- Have you considered grouping items into a matrix? (5 “sub-items” work well, definitely avoid matrices with >10 sub-items!)

- Are there contextual cues in your questions that may influence a respondent’s answer? (If yes, make sure you place them where people read them, i.e., in the sub-item rather than the stem).
Thank you!

jbertling@ets.org
jbertling@fordham.edu
Bertling.Jonas@gmail.com