

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

**10/14/2022**

**IDEE Workshop**

Paris School of Economics, Paris

Jonas Bertling, Ph.D.

Director, Educational Testing Service

Adjunct Professor, Fordham Graduate School of Education

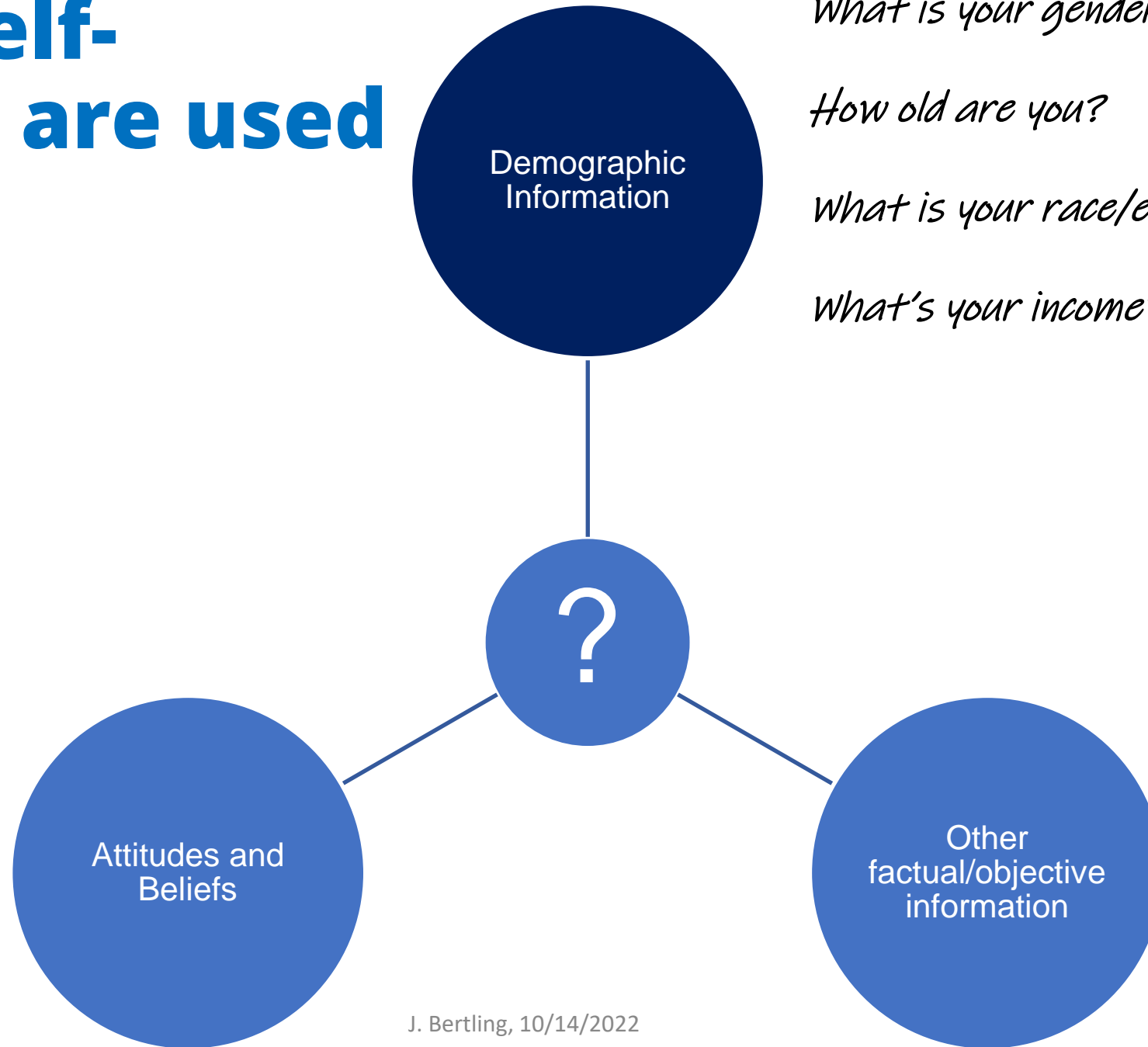
# Outline

- What self-report data are used for
- Reference groups and other challenges
- A case for more rigorous survey design principles
  - Question design, selection, interpretation

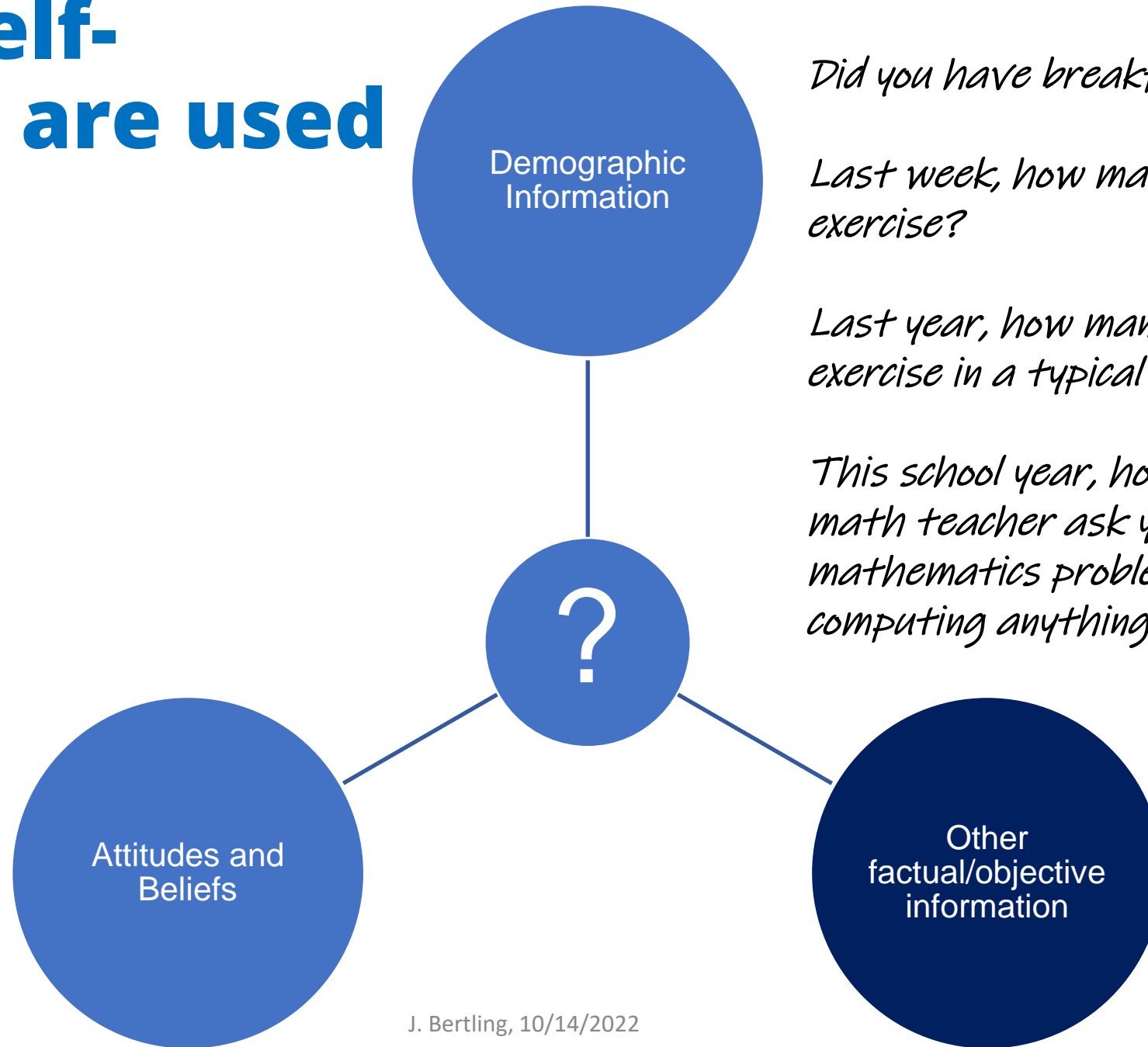
**Good questionnaire design may be an art, but it's definitely a science.**

**P.S. Thinking that good art does not require skill and discipline is an insult to the artist.**

# What self-reports are used for



# What self-reports are used for



*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?*

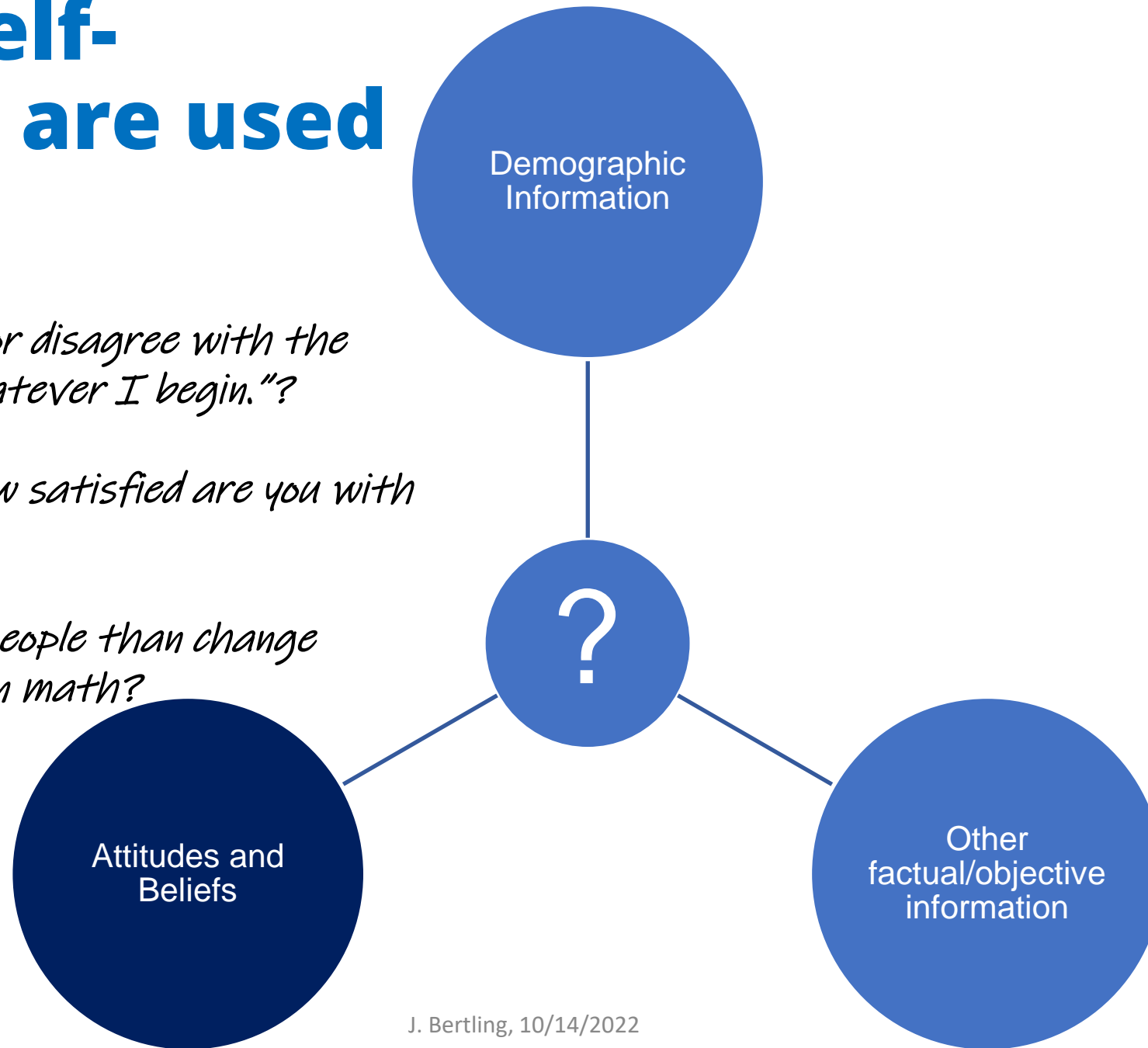
*This school year, how often did your math teacher ask you to solve mathematics problems without computing anything?*

# 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 than change whether they are good in math?*



# What the general public sees

PISA 2018

## 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

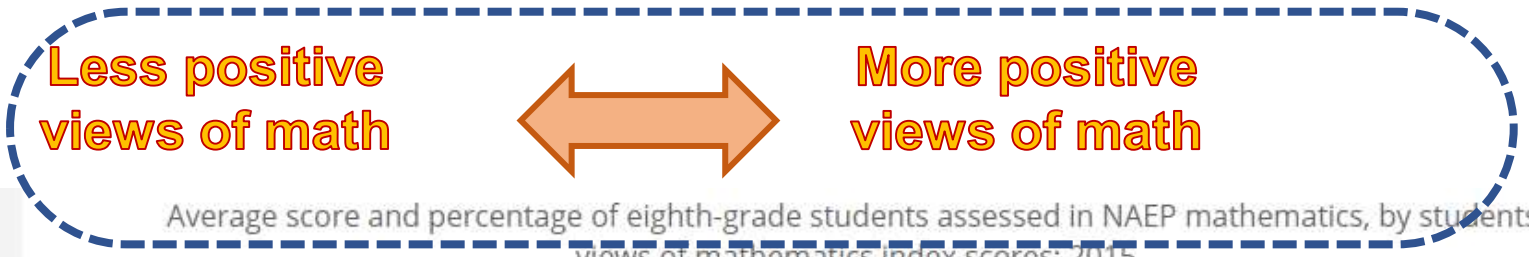
or

It is a good thing to help students who are being bullied defend themselves

**Bullying measure is based on student self-report questions.**

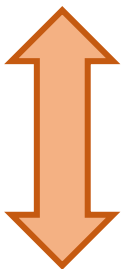
# What the general public sees (Cont'd)

NAEP 2015

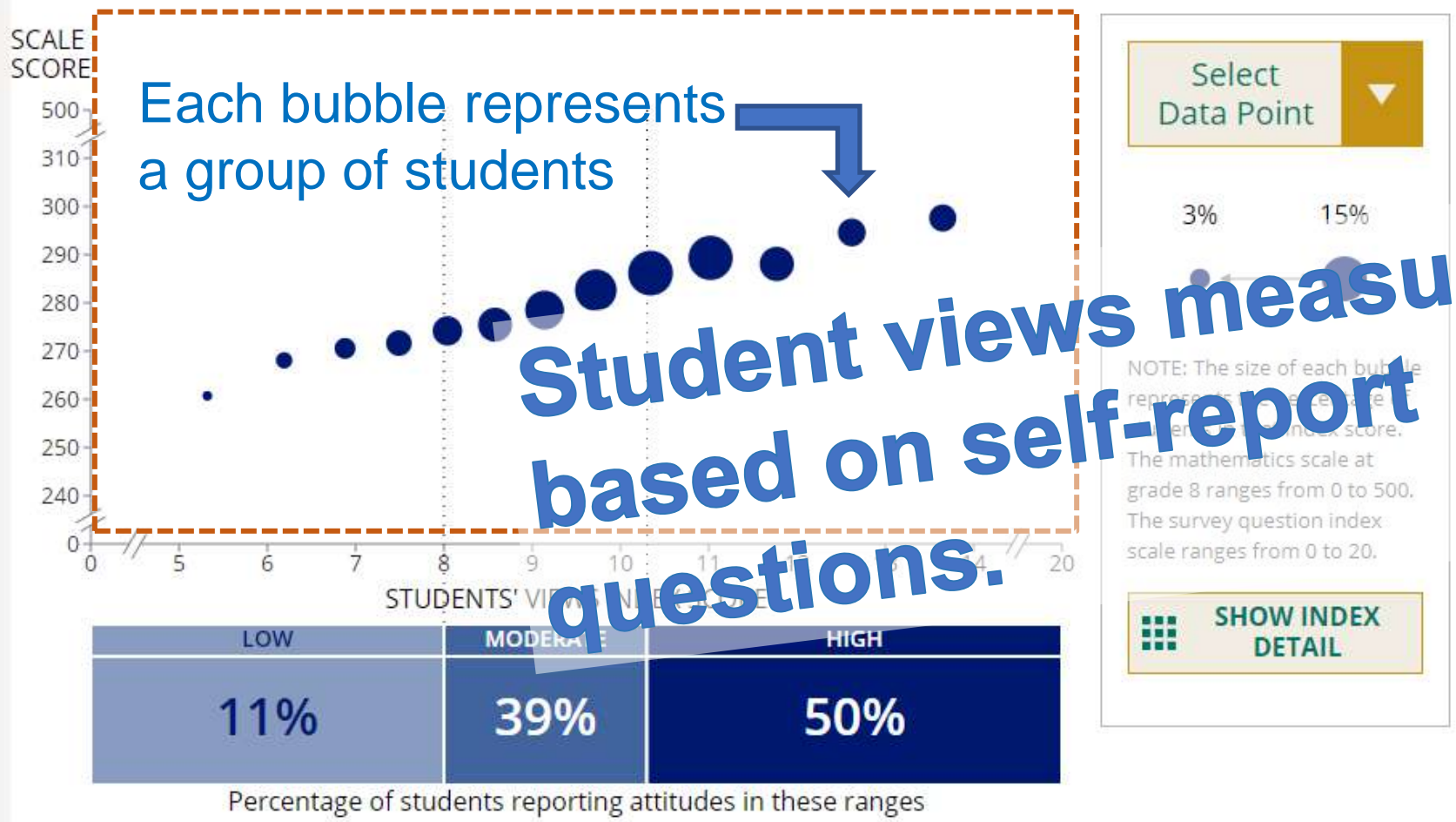


Average score and percentage of eighth-grade students assessed in NAEP mathematics, by students' views of mathematics index scores: 2015

Strong math performance



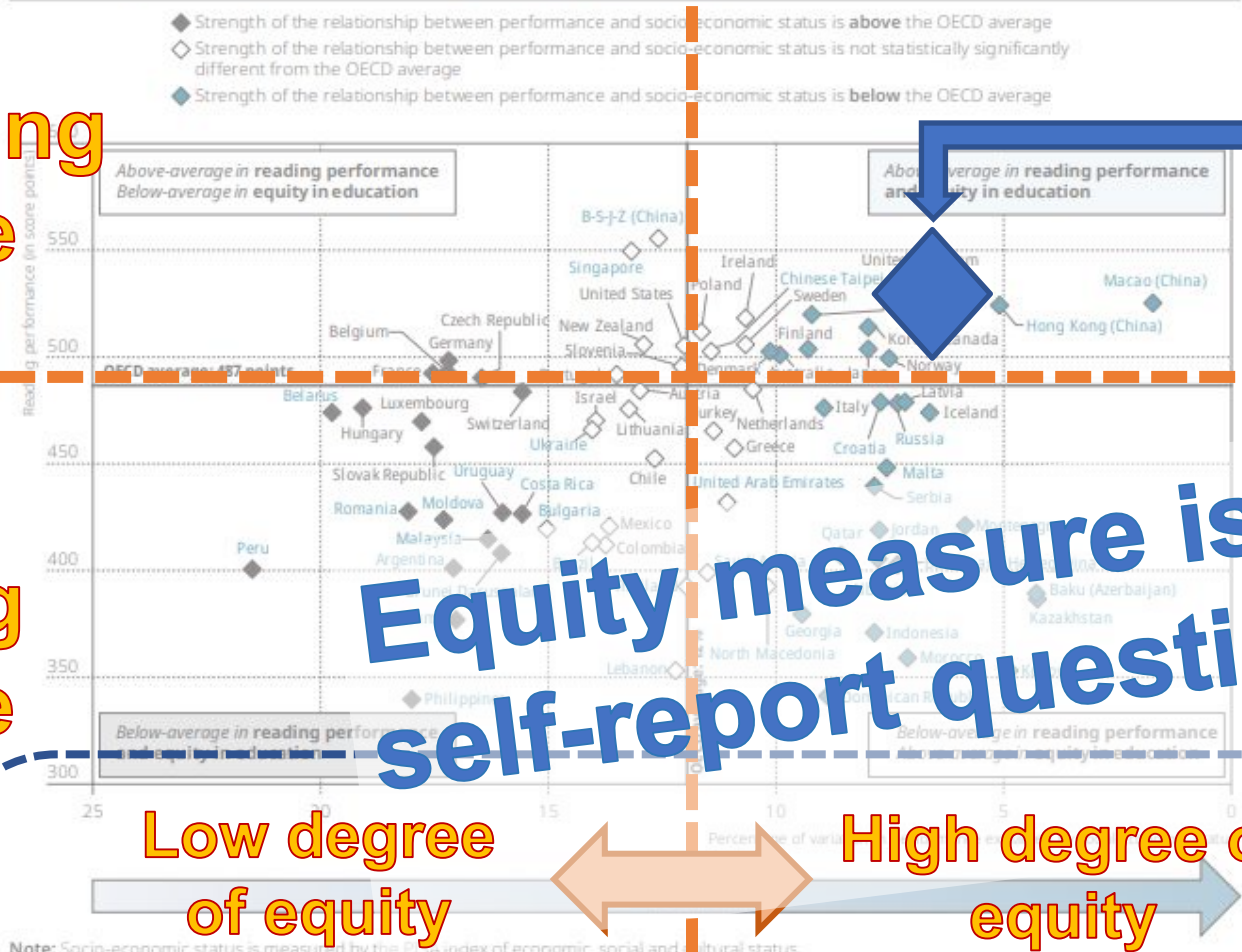
Poor math performance





# Strength of the socio-economic gradient and reading performance

Figure II.2.5 Strength of the socio-economic gradient and reading performance



Strong reading performance

Poor reading performance

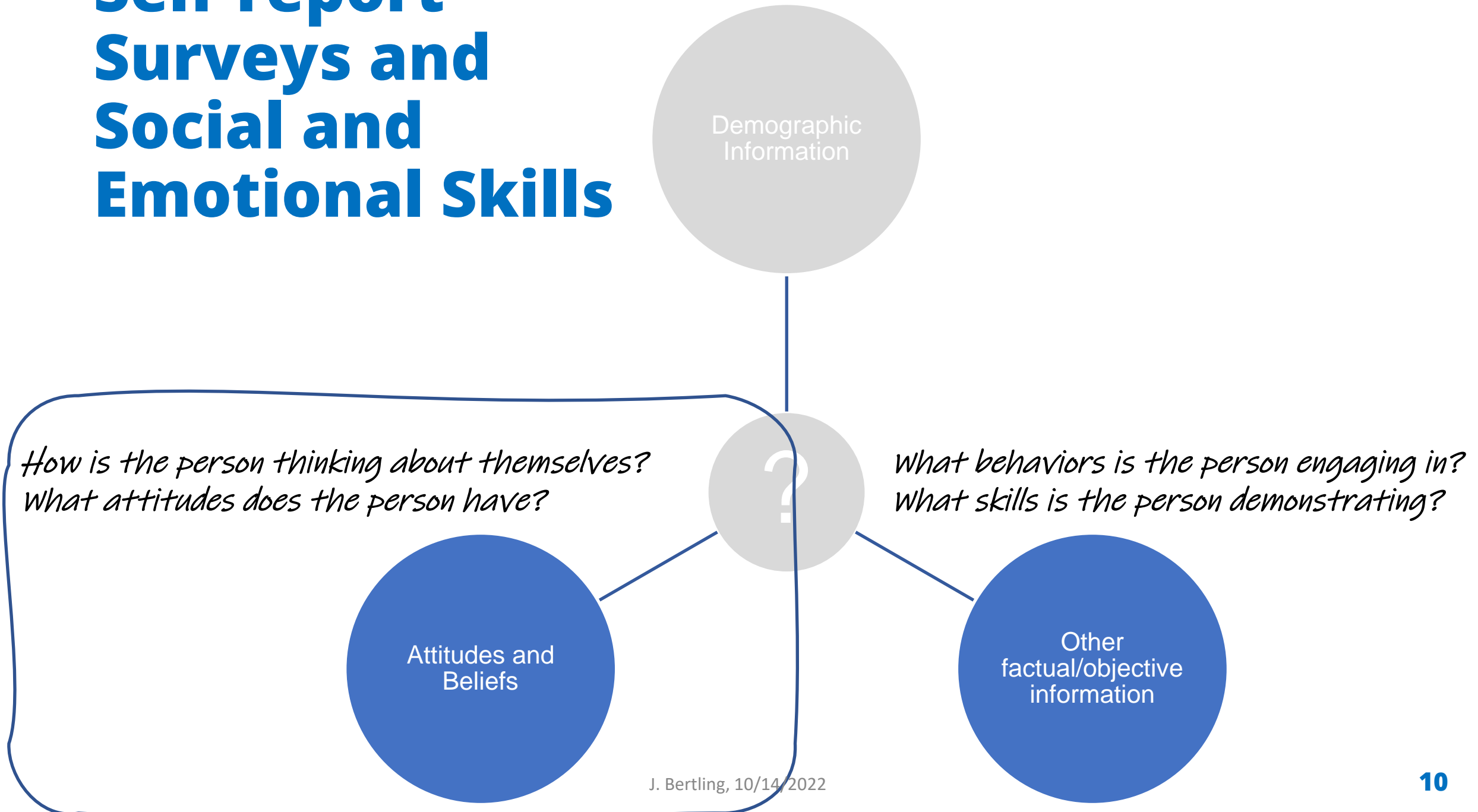
Each diamond is a country

Equity measure is based on self-report questions.0

Low degree of equity

High degree of equity

# Self-report Surveys and Social and Emotional Skills



# PISA 2022



# Reference groups and other challenges

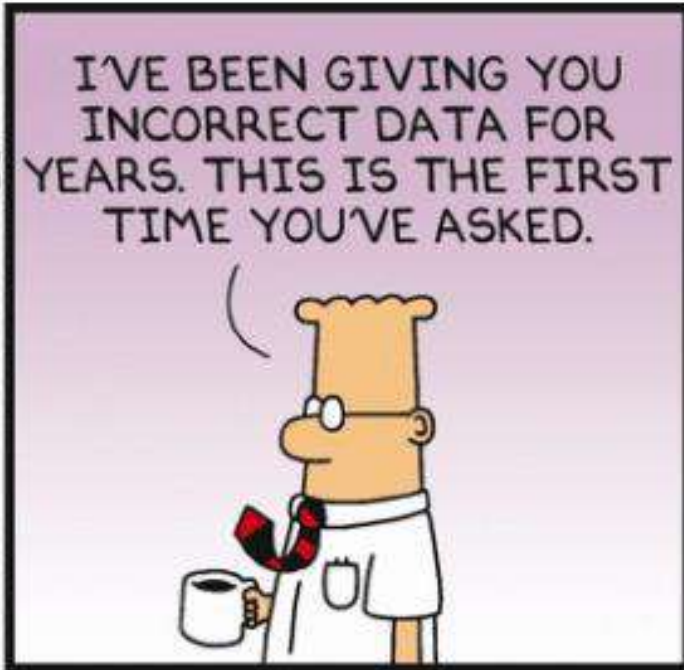
The indicator works differently across individuals from different groups

The indicator is not valid

The indicator is not reliable



Dilbert.com DilbertCartoonist@gmail.com



5-7-14 ©2014 Scott Adams, Inc. /Dist. by Universal Uclick



# It all starts with the item – There are many ways to end up with bad data

Demographics      Factual Information/  
behavioral reports      Attitudes/Beliefs

*Can respondents accurately calibrate their answer?*

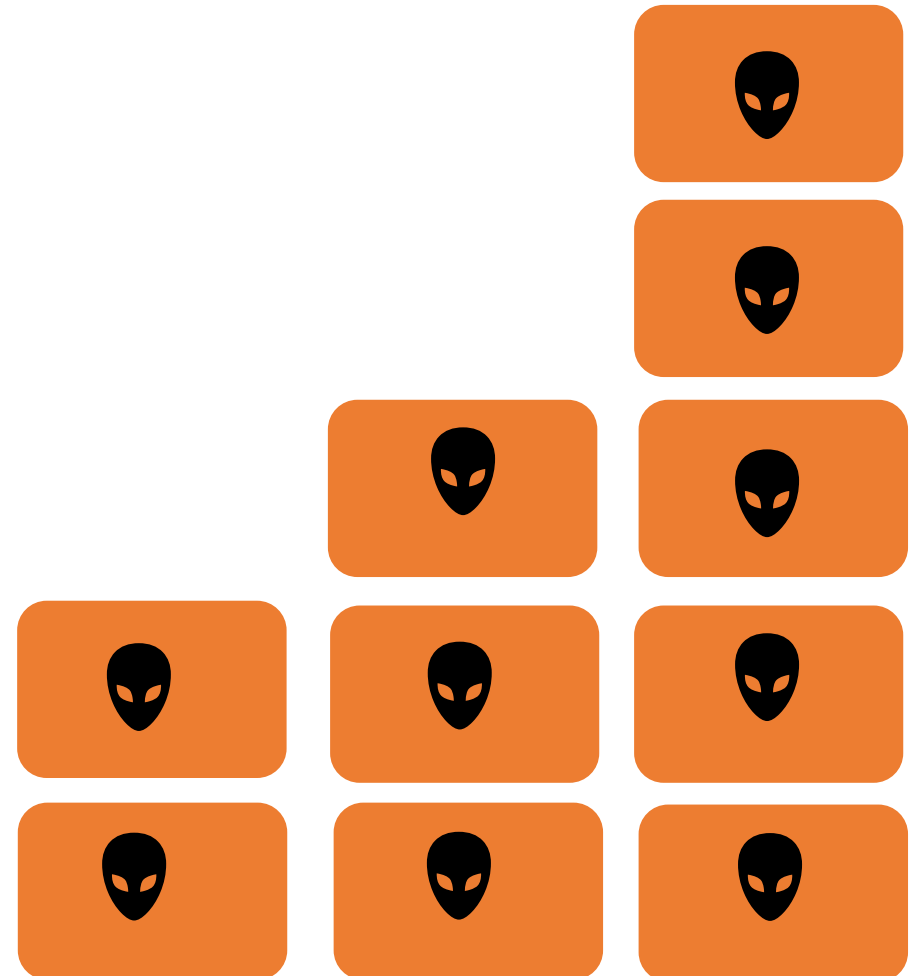
*Do respondents have necessary level of self-awareness?*

*Can respondents remember?*

*Are respondents willing to disclose accurate information?*

*Do all respondents understand the question in the same way?*

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.)*

		<i>Strongly agree</i>	<i>Agree</i>	<i>Disagree</i>	<i>Strongly disagree</i>
ST034Q01TA	I feel like an outsider (or left out of things) at school.	<input type="checkbox"/> <sub>01</sub>	<input checked="" type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q02TA	I make friends easily at school.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q03TA	I feel like I belong at school.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q04TA	I feel awkward and out of place in my school.	<input type="checkbox"/> <sub>01</sub>	<input checked="" type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q05TA	Other students seem to like me.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q06TA	I feel lonely at school.	<input type="checkbox"/> <sub>01</sub>	<input checked="" type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>



# Extreme Responding

ST034 **Thinking about your school: to what extent do you agree with the following statements?**

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

		<i>Strongly agree</i>	<i>Agree</i>	<i>Disagree</i>	<i>Strongly disagree</i>
ST034Q01TA	I feel like an outsider (or left out of things) at school.	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input checked="" type="checkbox"/> <sub>04</sub>
ST034Q02TA	I make friends easily at school.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q03TA	I feel like I belong at school.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q04TA	I feel awkward and out of place in my school.	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input checked="" type="checkbox"/> <sub>04</sub>
ST034Q05TA	Other students seem to like me.	<input checked="" type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q06TA	I feel lonely at school.	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input checked="" type="checkbox"/> <sub>04</sub>

# Reference Group Bias

“Big Fish Little Pond Effect”

*Strongly agree*

*Agree*

*Disagree*

*Strongly disagree*

ST034Q05TA	Other students seem to like me.	<input type="checkbox"/> <sub>01</sub>	<input checked="" type="checkbox"/> <sub>02</sub>	<input checked="" type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST034Q06TA	I feel lonely at school.	<input type="checkbox"/> <sub>01</sub>	<input checked="" type="checkbox"/> <sub>02</sub>	<input checked="" type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>

Not really, but more than Peter.

Yeah, but not as much as David.

Rarely, but I think more often than others

Most of the time, but compared to everyone else, not really

# 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

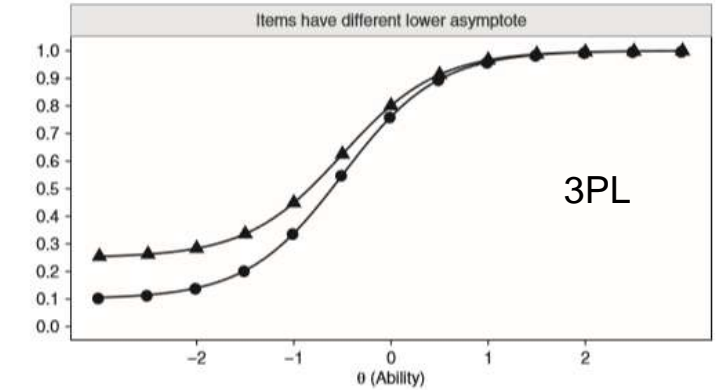
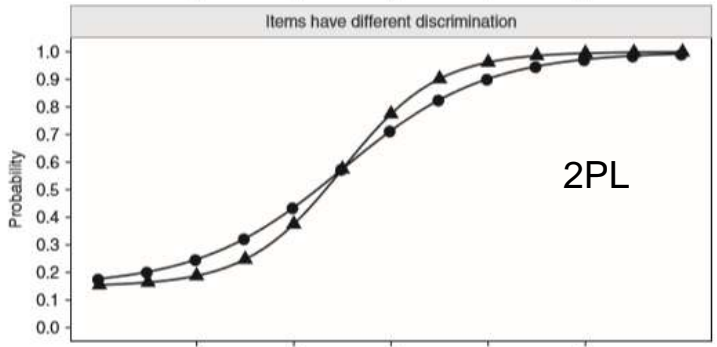
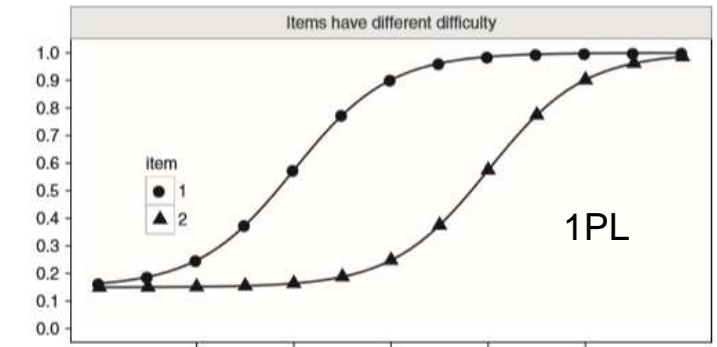
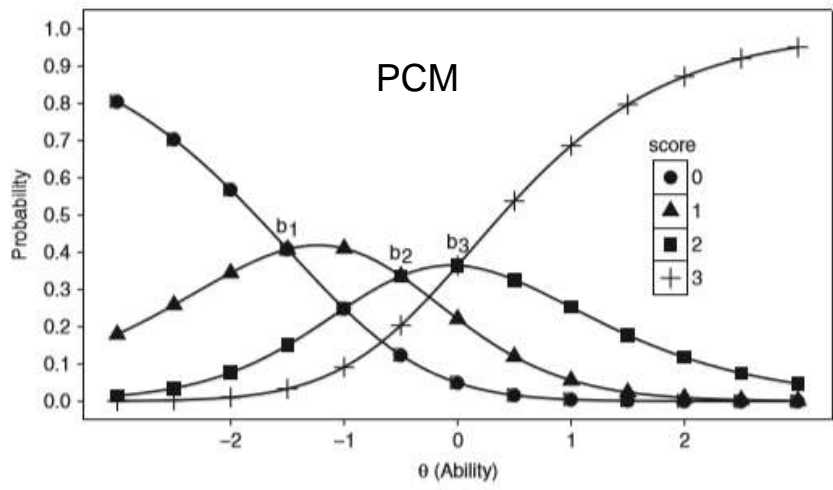
- **“Classical Test Theory”** describes the effects of measurement error on test scores
  - Error = not a mistake, but inconsistencies caused by random influences on test scores
- **Item Response Theory** models the probability of a correct response to an item (or agreement with a statement), conditional 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?





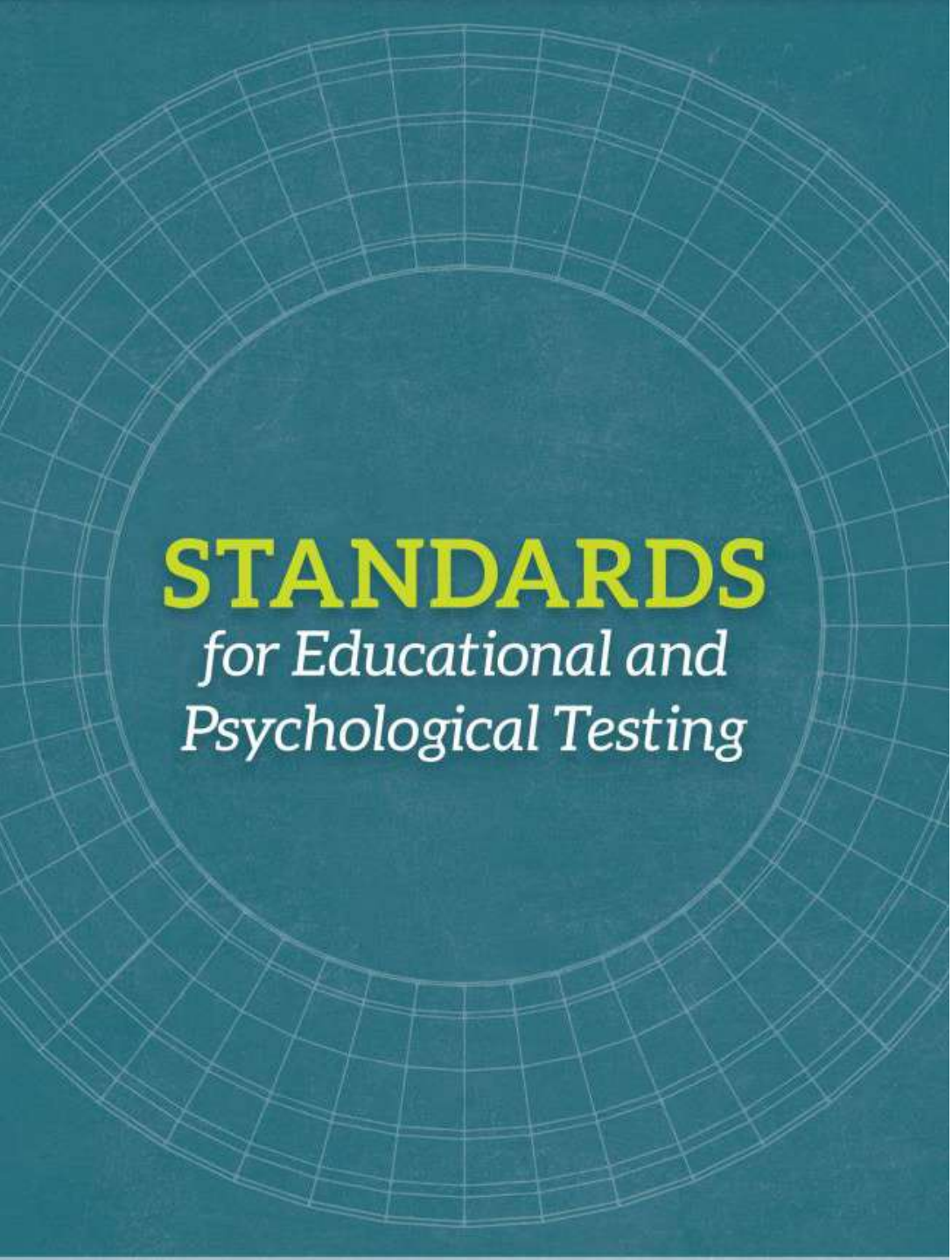
# Reliability and Validity in Everyday language

- **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?





# STANDARDS

*for Educational and  
Psychological Testing*

## “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?

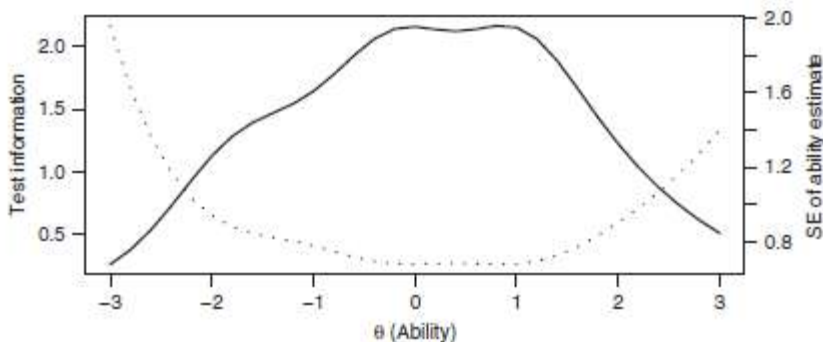
- Zumbo and Chan (2014): psychological scientists tend to report relatively little validity evidence and focus much more on other psychometric properties, most importantly reliability.
- Simplest explanation: providing reliability evidence is relatively easy, whereas providing validity evidence is very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

# Common methods to assess Reliability

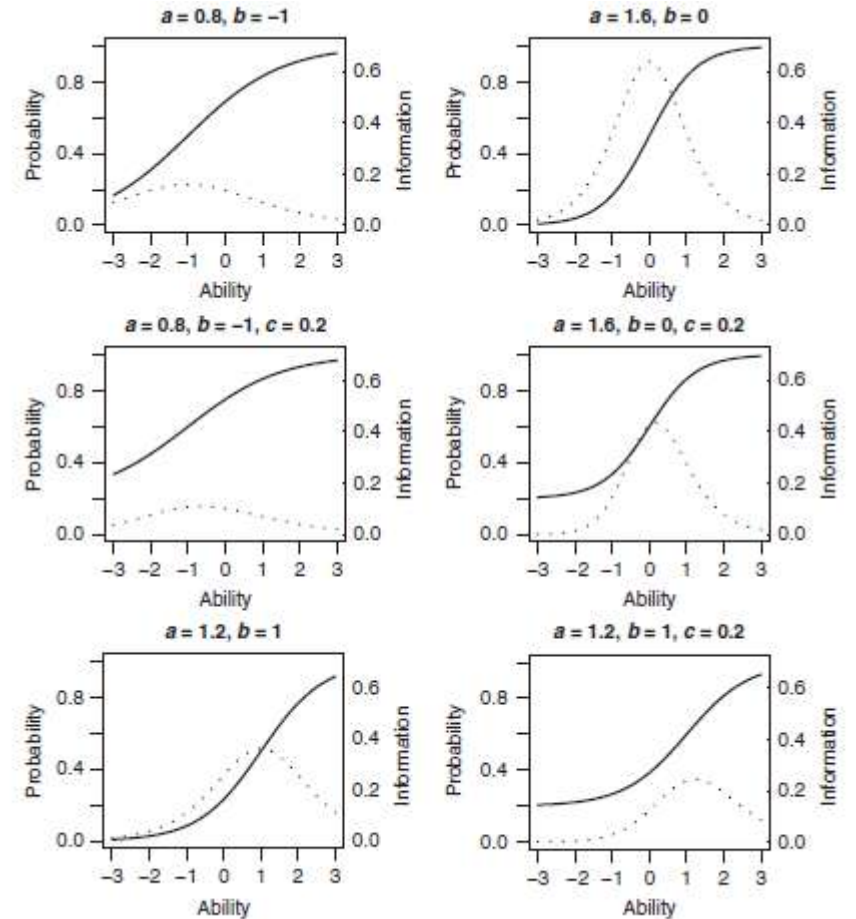
- **Internal consistency** (Cronbach's Alpha): average correlation between all items of the test
- **Split-half** (coefficient of stability): correlation between two "halves" of the test
- **Test-retest** (coefficient of stability): correlation between scores on the same test collected at two different 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:



$$1/\sqrt{I(\theta)}$$



# 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

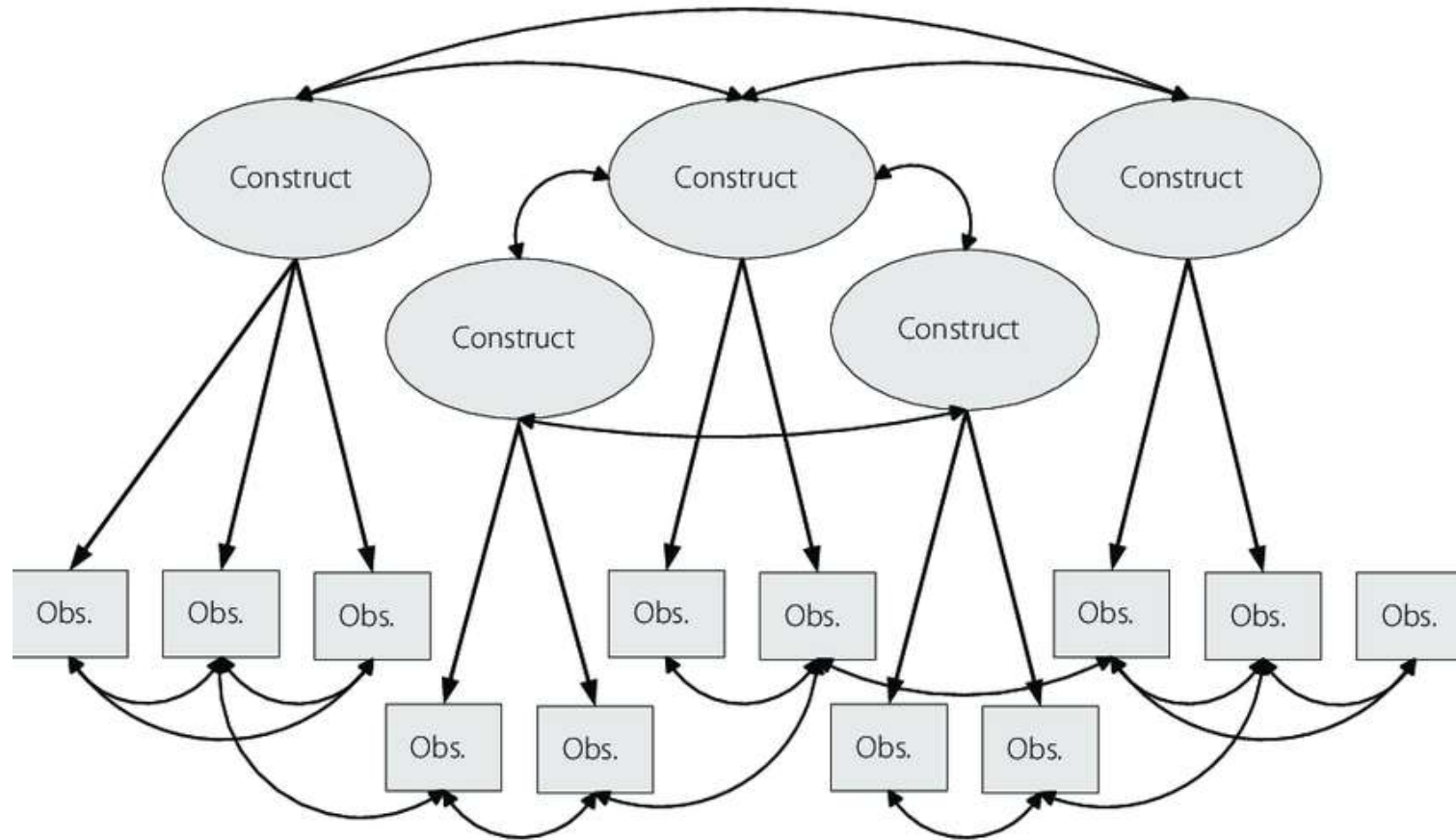
- Based on BFI-44

<b>Big Five Dimensions</b>	<b>Facet (and correlated trait adjective)</b>
Extraversion vs. introversion	Gregariousness (sociable) Assertiveness (forceful) Activity (energetic) Excitement-seeking (adventurous) Positive emotions (enthusiastic) 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)

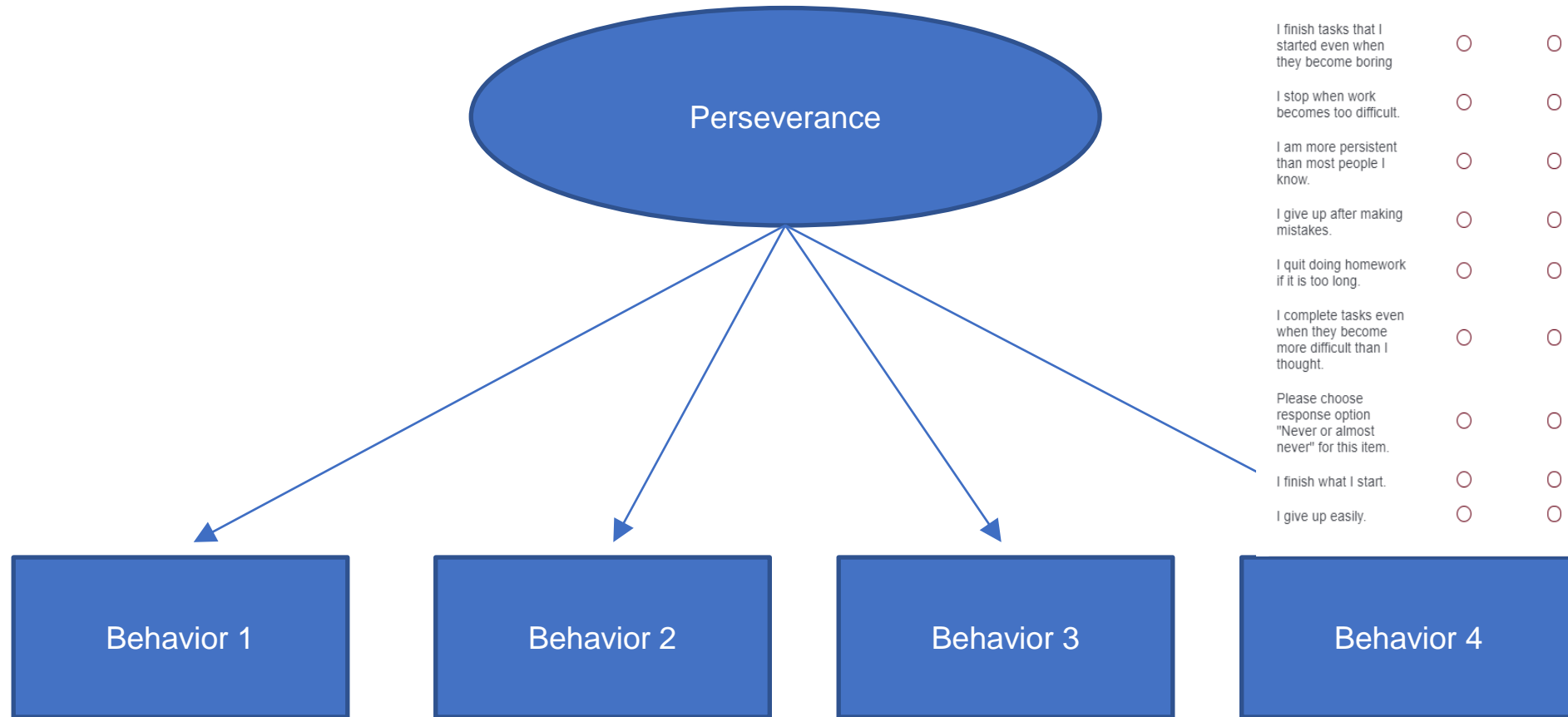
<https://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Personality-BigFiveInventory.pdf>



# Nomological Nets



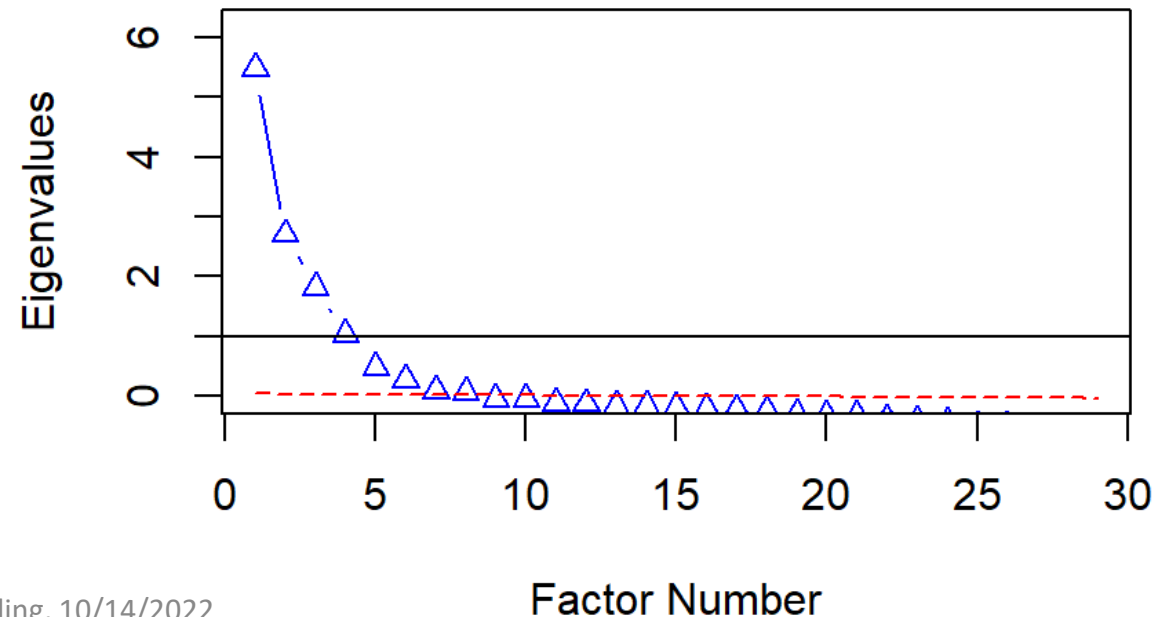
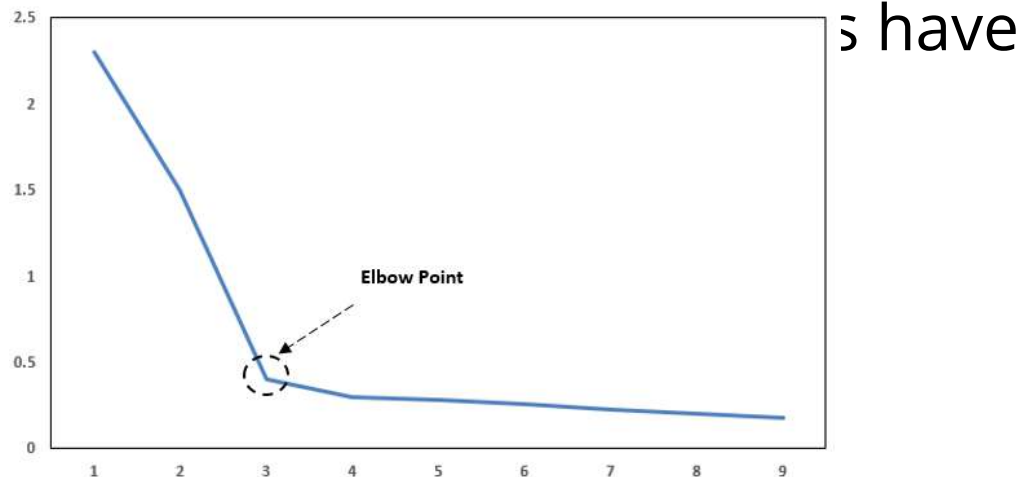
How often do you think, feel, or act in the following ways?



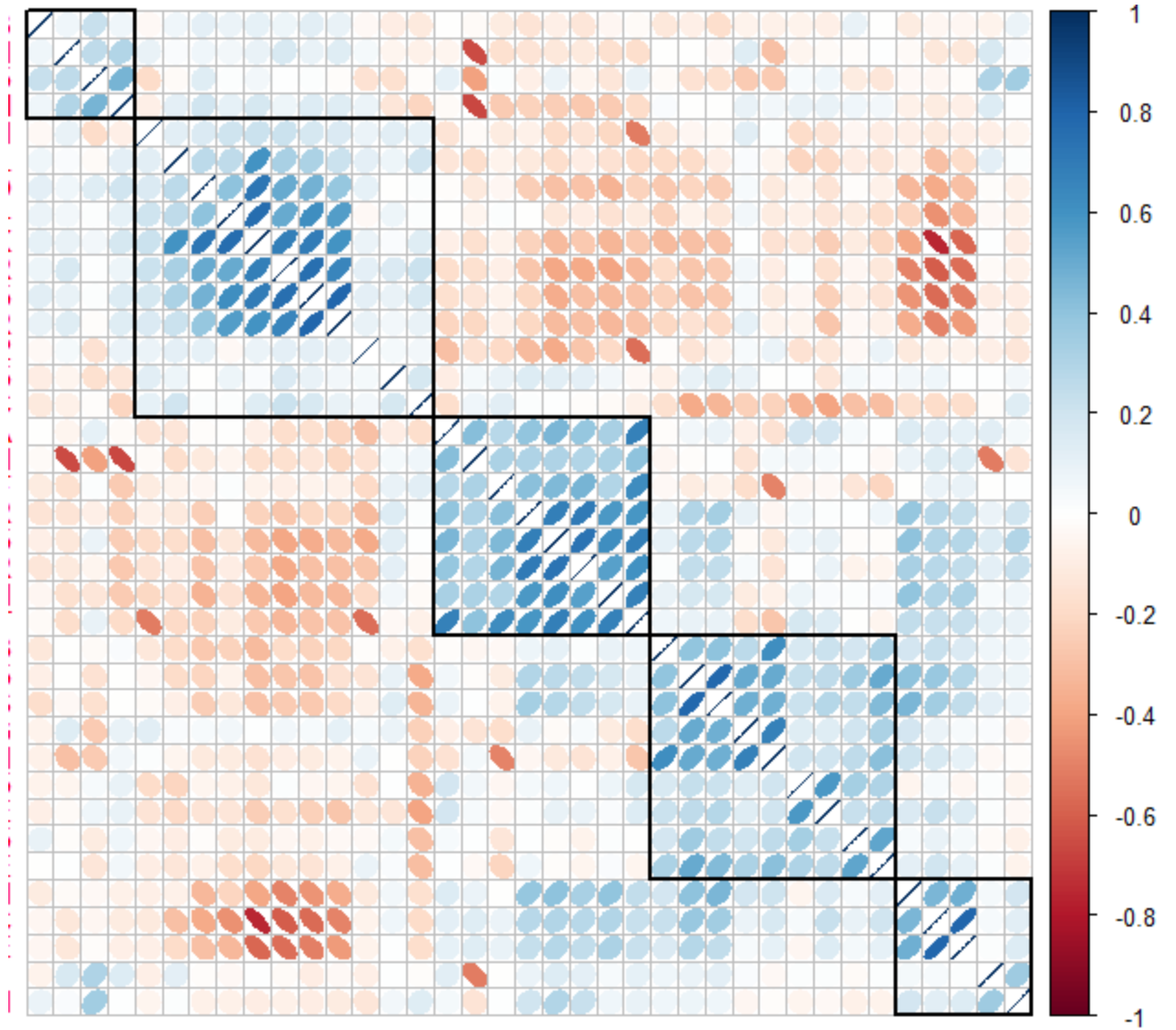
	Never or almost never	Less than half of the time	About half of the time	More than half of the time	All or almost all of the time
I apply additional effort when work becomes challenging.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I keep working on a task until it is finished.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I finish tasks that I started even when they become boring	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I stop when work becomes too difficult.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am more persistent than most people I know.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I give up after making mistakes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I quit doing homework if it is too long.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I complete tasks even when they become more difficult than I thought.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Please choose response option "Never or almost never" for this item.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I finish what I start.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I give up easily.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

# 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

Pattern Matrix <sup>a</sup>					
Factor					
1	2	3	4	5	
		0.428		-0.301	
0.207			-0.404	-0.235	
	-0.391			0.221	
	0.355	-0.313		0.260	
			0.357	0.265	
		0.543		0.272	
	-0.314			0.282	
0.493				0.287	
	-0.460			0.296	
0.256				0.304	
	-0.234	0.277		0.385	
	-0.223			0.458	
		0.255		0.462	
				0.483	
				0.490	
0.495	-0.203		-0.266		
	0.634		-0.246		
-0.283		0.392	-0.237		
0.667			-0.231		
		0.237	0.219		
0.262			0.450		
			0.455		
-0.214			0.456		
			0.520		
		0.237	0.528		
		-0.217	0.548		
		-0.714			
	0.338	-0.592			
	0.605	0.369			
0.321		0.473			

- 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.**

# Short Grit Scale

Please respond to the following 8 items. Be honest - there are no right or wrong answers.

	Very much like me	Mostly like me	Somewhat like me	Not much like me	Not like me at all
New ideas and projects sometimes distract me from previous ones.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Setbacks don't discourage me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have been obsessed with a certain idea or project for a short time but later lost interest.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am a hard worker	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I often set a goal but later choose to pursue a different one.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
i have difficulty maintaining my focus on projects that take more than a few months to complete.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I finish whatever I begin.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am diligent.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

“Duckworth, Peterson, Matthews, and Kelly (2007) introduced the construct of grit, defined as **trait-level perseverance and passion for long-term goals.**” (p. 166)

*Journal of Personality Assessment*, 91(2), 166–174, 2009  
 Copyright © Taylor & Francis Group, LLC  
 ISSN: 0022-3891 print / 1532-7752 online  
 DOI: 10.1080/00223890802634290



## Development and Validation of the Short Grit Scale (Grit-S)

ANGELA LEE DUCKWORTH AND PATRICK D. QUINN

*Department of Psychology, University of Pennsylvania*

In this article, we introduce brief self-report and informant-report versions of the Grit Scale, which measures trait-level perseverance and passion for long-term goals. The Short Grit Scale (Grit-S) retains the 2-factor structure of the original Grit Scale (Duckworth, Peterson, Matthews, & Kelly, 2007) with 4 fewer items and improved psychometric properties. We present evidence for the Grit-S's internal consistency, test-retest stability, consensual validity with informant-report versions, and predictive validity. Among adults, the Grit-S was associated with educational attainment and fewer career changes. Among adolescents, the Grit-S longitudinally predicted GPA and, inversely, hours watching television. Among cadets at the United States Military Academy, West Point, the Grit-S predicted retention. Among Scripps National Spelling Bee competitors, the Grit-S predicted final round attained, a relationship mediated by lifetime spelling practice.

Perseverance is more often studied as an outcome than as a predictor. For example, perseverance in difficult or impossible tasks has served as the dependent variable in studies of optimistic attribution style, self-efficacy, goal orientation, and depletion of self-control resources (see, e.g., Bandura, 1977; Baumeister, Bratslavsky, Muraven, & Tice, 1998; Elliott & Dweck, 1988; Muraven, Tice, & Baumeister, 1998; Seligman & Schulman, 1986). However, the study of perseverance as a predictor, in particular as a stable individual difference, was of keen interest to psychologists in the first half of the 20th century. In a review of the existing literature of his day, Ryans (1939) concluded that “the existence of a general trait of persistence, which permeates all behavior of the organism, has not been established, though evidence both for and against such an assumption has been revealed” (p. 737). Very recently, positive psychology has renewed interest in the empirical study of character in general and in the trait of perseverance in particular (Peterson & Seligman, 2004).

Duckworth, Peterson, Matthews, and Kelly (2007) introduced the construct of *grit*, defined as trait-level perseverance and passion for long-term goals, and showed that grit predicted achievement in challenging domains over and beyond measures of talent. For instance, at the U.S. Military Academy, West Point, cadets higher in grit were less likely to drop out than their less

There is ample evidence that the moderate challenge incentive is crucial for individuals high in *n* Achievement; they will work harder when this incentive is present than when it is not present; that is, when tasks are too easy or too hard [italics added]. (p. 814)

Duckworth et al. (2007) identified a two-factor structure for the original 12-item self-report measure of grit (Grit-O). This structure was consistent with the theory of grit as a compound trait comprising stamina in dimensions of interest and effort. However, the differential predictive validity of these two factors for various outcomes was not explored. Duckworth et al. did not examine whether either factor predicted outcomes better than did the other. Moreover, the model fit of the Grit-O (comparative fit index [CFI]<sup>1</sup> = .83; root mean square error of approximation [RMSEA]<sup>2</sup> = .11) suggested room for improvement.

### THIS RESEARCH

We undertook this investigation to validate a more efficient measure of grit. In Study 1, we identified items for the Short Grit Scale (Grit-S) with the best overall predictive validity across four samples originally presented in Duckworth et al. (2007). In Study 2, we used confirmatory factor analysis to test the two-factor structure of the Grit-S in a novel Internet sample of adults, compared the relationships between the Grit-S and Grit-O and

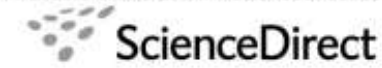
# Conscientiousness via BFI-10

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

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
... is reserved	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... is generally trusting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... tends to be lazy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... is relaxed, handles stress well	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... has few artistic interests	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... is outgoing, sociable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... tends to find fault with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... does a thorough job	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... gets nervous easily	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... has an active imagination	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Journal of Research in Personality 41 (2007) 203–212

JOURNAL OF  
RESEARCH IN  
PERSONALITY

[www.elsevier.com/locate/jrp](http://www.elsevier.com/locate/jrp)

Brief report

## Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German <sup>☆</sup>

Beatrice Rammstedt <sup>a,\*</sup>, Oliver P. John <sup>b</sup>

<sup>a</sup> Center for Survey Research and Methodologies (ZUMA), P.O. Box 12 21 55, D-68072 Mannheim, Germany

<sup>b</sup> Department of Psychology, University of California, Berkeley MC 1650, Berkeley, CA 94720-1650, USA

Available online 3 April 2006

### 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	Competence (efficient) Order (organized) Dutifulness (not careless) Achievement striving (thorough) Self-discipline (not lazy) Deliberation (not impulsive)
---	--

- \_\_\_ 1. Is talkative
- \_\_\_ 2. Tends to find fault with others
- \_\_\_ 3. Does a thorough job**
- \_\_\_ 4. Is depressed, blue
- \_\_\_ 5. Is original, comes up with new ideas
- \_\_\_ 6. Is reserved
- \_\_\_ 7. Is helpful and unselfish with others
- \_\_\_ 8. Can be somewhat careless**
- \_\_\_ 9. Is relaxed, handles stress well
- \_\_\_ 10. Is curious about many different things
- \_\_\_ 11. Is full of energy
- \_\_\_ 12. Starts quarrels with others
- \_\_\_ 13. Is a reliable worker**
- \_\_\_ 14. Can be tense
- \_\_\_ 15. Is ingenious, a deep thinker
- \_\_\_ 16. Generates a lot of enthusiasm
- \_\_\_ 17. Has a forgiving nature
- \_\_\_ 18. Tends to be disorganized**
- \_\_\_ 19. Worries a lot
- \_\_\_ 23. Tends to be lazy**
- \_\_\_ 24. Is emotionally stable, not easily upset
- \_\_\_ 25. Is inventive
- \_\_\_ 26. Has an assertive personality
- \_\_\_ 27. Can be cold and aloof
- \_\_\_ 28. Perseveres until the task is finished**
- \_\_\_ 29. Can be moody
- \_\_\_ 30. Values artistic, aesthetic experiences
- \_\_\_ 31. Is sometimes shy, inhibited
- \_\_\_ 32. Is considerate and kind to almost everyone
- \_\_\_ 33. Does things efficiently**
- \_\_\_ 34. Remains calm in tense situations
- \_\_\_ 35. Prefers work that is routine
- \_\_\_ 36. Is outgoing, sociable
- \_\_\_ 37. Is sometimes rude to others
- \_\_\_ 38. Makes plans and follows through with them**
- \_\_\_ 39. Gets nervous easily
- \_\_\_ 40. Likes to reflect, play with ideas
- \_\_\_ 41. Has few artistic interests

- \_\_\_ 20. Has an active imagination
- \_\_\_ 21. Tends to be quiet
- \_\_\_ 22. Is generally trusting
- \_\_\_ 42. Likes to cooperate with others
- \_\_\_ 43. Is easily distracted**
- \_\_\_ 44. Is sophisticated in art, music, or literature

### 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

# Survey Questionnaires Development Phases in Large-Scale Assessments



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?*

Can be measured with one question

## Not directly observable

- *How perseverant is student X?*
- *What is student X's socio-economic status?*

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

	# Items	Mean
BELONG	3	0.559
	4	0.634
	5	0.687
COGACT	3	0.584
	4	0.655
	5	0.705
MATHWKETH	3	0.682
	4	0.744
	5	0.785
MATHEFF	3	0.627
	4	0.694
	5	0.742
MATHBEH	3	0.471
	4	0.540
	5	0.593
FAMCON	3	0.476
	4	0.549
	5	0.605

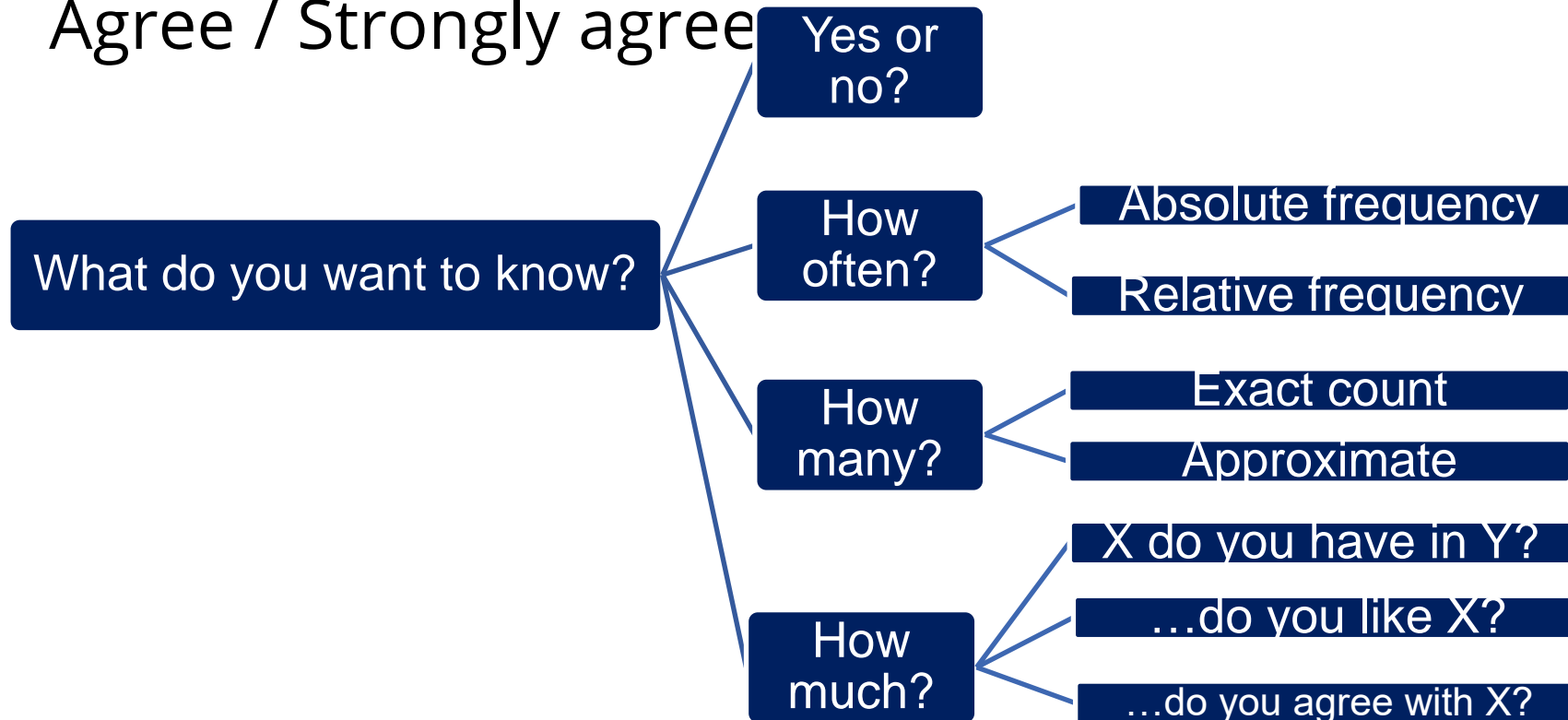
From: Bertling & Weeks (2020)

**How can you choose the most appropriate response options?**

# Don't default to "agree-disagree" options.

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

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



**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

**Use labels for all scale points**  
**Cover the entire range of possible answers**

**Don't use too many or too few response options**

**Avoid vague terms like "sometimes", "often", "rarely"**  
**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?

- A I definitely can't.
- B I probably can't.
- C I probably can.
- D I definitely can.

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

- A I definitely can't.
- B I probably can't.
- C I probably can.
- D I definitely can.

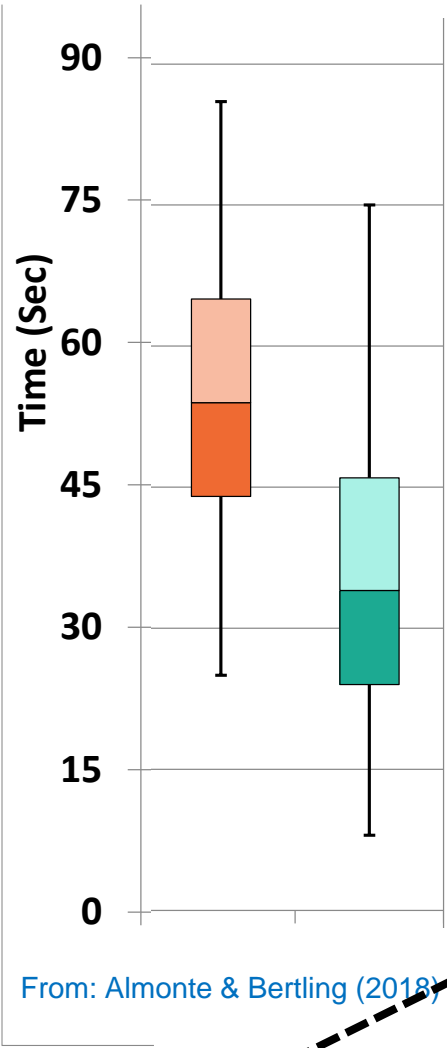
VS.

Do you think you would be able to do each of the following? Select **one** answer choice on each row.

	I definitely can't.	I probably can't.	I probably can.	I definitely can.
a. Write sentences and paragraphs using a computer	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D
b. Edit text using a computer	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D
c. Use a touchscreen on a computer, tablet, or smartphone	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D
d. Look up the meaning of a word using a computer	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D
e. Draw a picture using a computer	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D

Same questions in two different formats

Findings from NAEP study with 4<sup>th</sup> graders:



- Matrix questions take less time to answer.  
- No notable differences in statistical data quality

But:

- Don't make matrices too long.
- Don't combine questions that cannot be combined.
- Make sure the stem matches the items.

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

34. How much does each of the following statements describe you? Select **one** answer choice on each row.

VH717301

	Not at all like me	A little bit like me	Somewhat like me	Quite a bit like me	Exactly like me	
a. I want to learn as much as possible about <b>geography</b> in my class.	(A)	(B)	(C)	(D)	(E)	VH717302
b. I want to master a lot of new <b>geography</b> skills in my class.	(A)	(B)	(C)	(D)	(E)	VH717303
c. I want to become a better <b>geography</b> student this year.	(A)	(B)	(C)	(D)	(E)	VH717305
d. I want to understand as much as I can about <b>geography</b> in my class.	(A)	(B)	(C)	(D)	(E)	VH717306

VS.

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

a. I want to learn as much as possible <b>_____</b> in my class.
b. I want to master a lot of new <b>_____</b> skills in my class.
c. I want to become a better <b>_____</b> student this year.
d. I want to understand as much as I can <b>_____</b> in my class.

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

Qureshi, Gill, Alegre, & Bertling (2018)

Average difference "stem-context" - "item-context"	Performance Goals	Civics		Geography		U.S. History	
		Not at all like me	Exactly like me	Not at all like me	Exactly like me	Not at all like me	Exactly like me
a. I want much about class.		-3.03%	6.92%	-3.68%	6.11%	-2.12%	4.41%
b. I want of new skills in	Mastery Goals	-7.58%	22.63%	-7.18%	19.43%	-4.80%	13.56%
c. I want to better ge student t							
d. I want to as much a geography							

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**

ST290 How confident do you feel about having to do the following mathematics tasks?  
(Please select one response in each row.)

	Not at all confident	Not very confident	Confident	Very confident
ST290Q01JA Working out from a <train timetable> how long it would take to get from one place to another	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q02WA Calculating how much more expensive a computer would be after adding tax	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q03WA Calculating how many square metres of tiles you need to cover a floor	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q04WA Understanding scientific tables presented in an article	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q05WA Solving an equation like $6x^2 + 5 = 29$	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q06WA Finding the actual distance between two places on a map with a 1:10,000 scale	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q07WA Solving an equation like $2(x+3) = (x+3)(x-3)$	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q08WA Calculating the power consumption of an electronic appliance per week	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q09WA Solving an equation like $3x+5=17$	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>

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

ST290 How confident do you feel about having to do the following mathematics tasks?  
(Please select one response in each row.)

	Not at all confident	Not very confident	Confident	Very confident
ST290Q01JA Working out from a <train timetable> how long it would take to get from one place to another	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q02WA Calculating how much more expensive a computer would be after adding tax	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q06WA Finding the actual distance between two places on a map with a 1:10,000 scale	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q07WA Solving an equation like $2(x+3) = (x+3)(x-3)$	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>
ST290Q08WA Calculating the power consumption of an electronic appliance per week	<input type="checkbox"/> <sub>01</sub>	<input type="checkbox"/> <sub>02</sub>	<input type="checkbox"/> <sub>03</sub>	<input type="checkbox"/> <sub>04</sub>

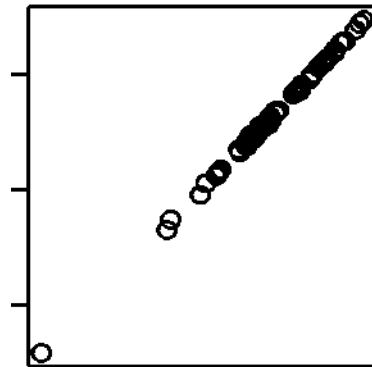
# Why >five items per construct?

**Matrix Sampling:**  
Random selection of 5 items per student

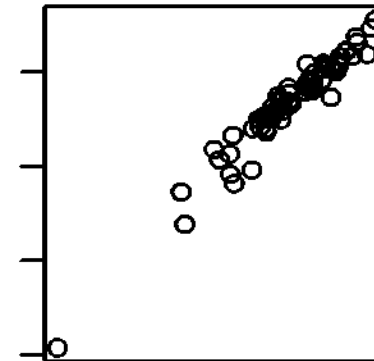
**Fixed Scale Shortening:**  
Administration of “5 best” items to every student

**PISA 2012  
Math Self-  
Efficacy**

Matrix-  
Sampled  
Scale

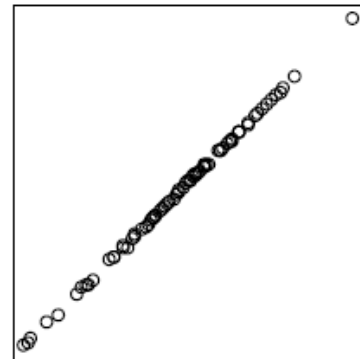


Short Scale

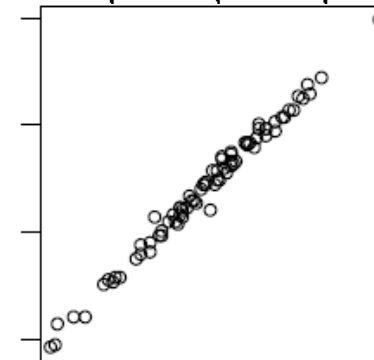


**PISA 2015  
Sense of  
Belonging**

Matrix-  
Sampled  
Scale



Short Scale



All items

J. Bertling, 10/14/2022

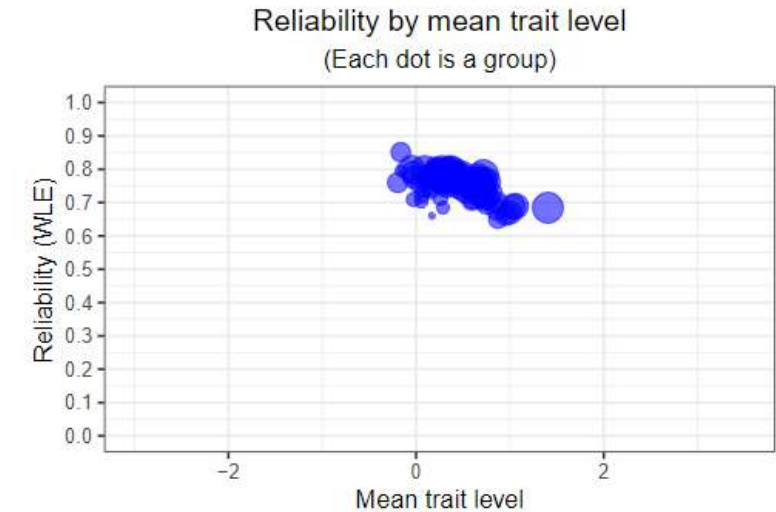
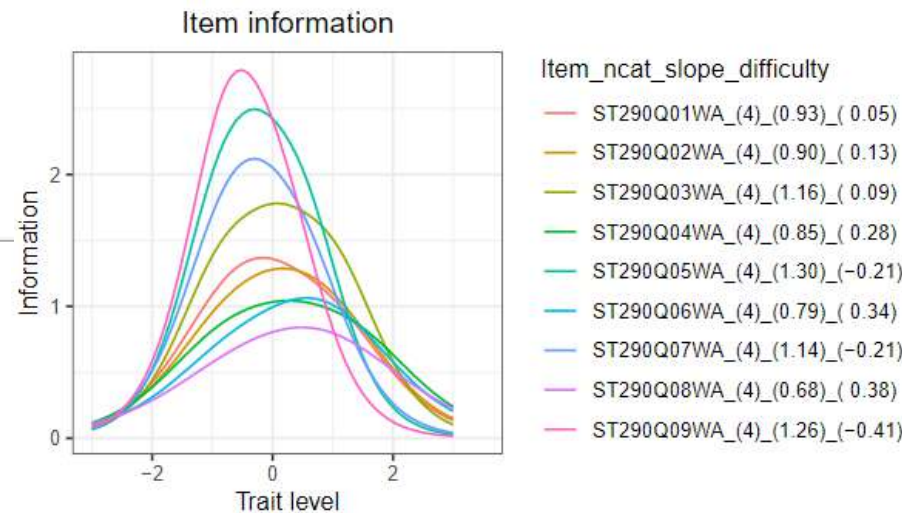
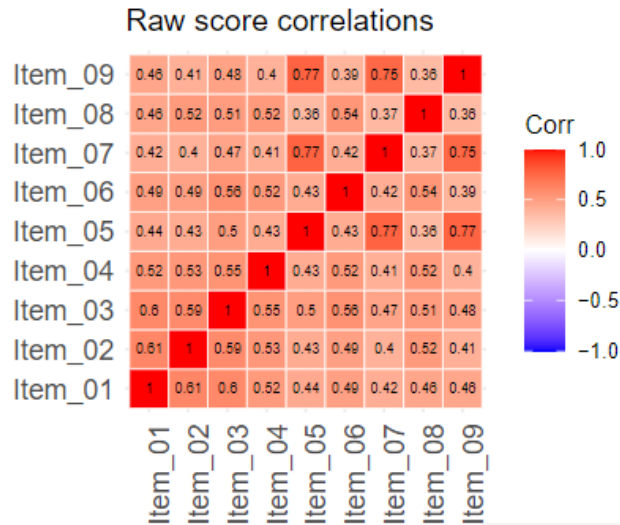
All items

# Feasibility confirmed in PISA 2022 FT

## Example: Self-efficacy

Item	Item_ID	Ncat	Slope	Difficulty	Step1	Step2	Step3
01	ST290Q01WA	4	0.93	0.05	1.08	0.08	-1.15
02	ST290Q02WA	4	0.90	0.13	1.14	-0.01	-1.13
03	ST290Q03WA	4	1.16	0.09	1.09	0.00	-1.09
04	ST290Q04WA	4	0.85	0.28	1.34	0.01	-1.35
05	ST290Q05WA	4	1.30	-0.21	0.85	0.02	-0.87
06	ST290Q06WA	4	0.79	0.34	1.24	-0.11	-1.13
07	ST290Q07WA	4	1.14	-0.21	0.85	0.03	-0.89
08	ST290Q08WA	4	0.68	0.38	1.30	-0.06	-1.24
09	ST290Q09WA	4	1.26	-0.41	0.69	0.06	-0.74

	Values
N_items	9
Median_responses_person	5
Median_Cronbach_alpha	0.83
Median_theta_reliability	0.76
Scale_scores	WLE





# Summing it up

# There are many ways to end up with bad data

*Can respondents accurately calibrate their answer?*

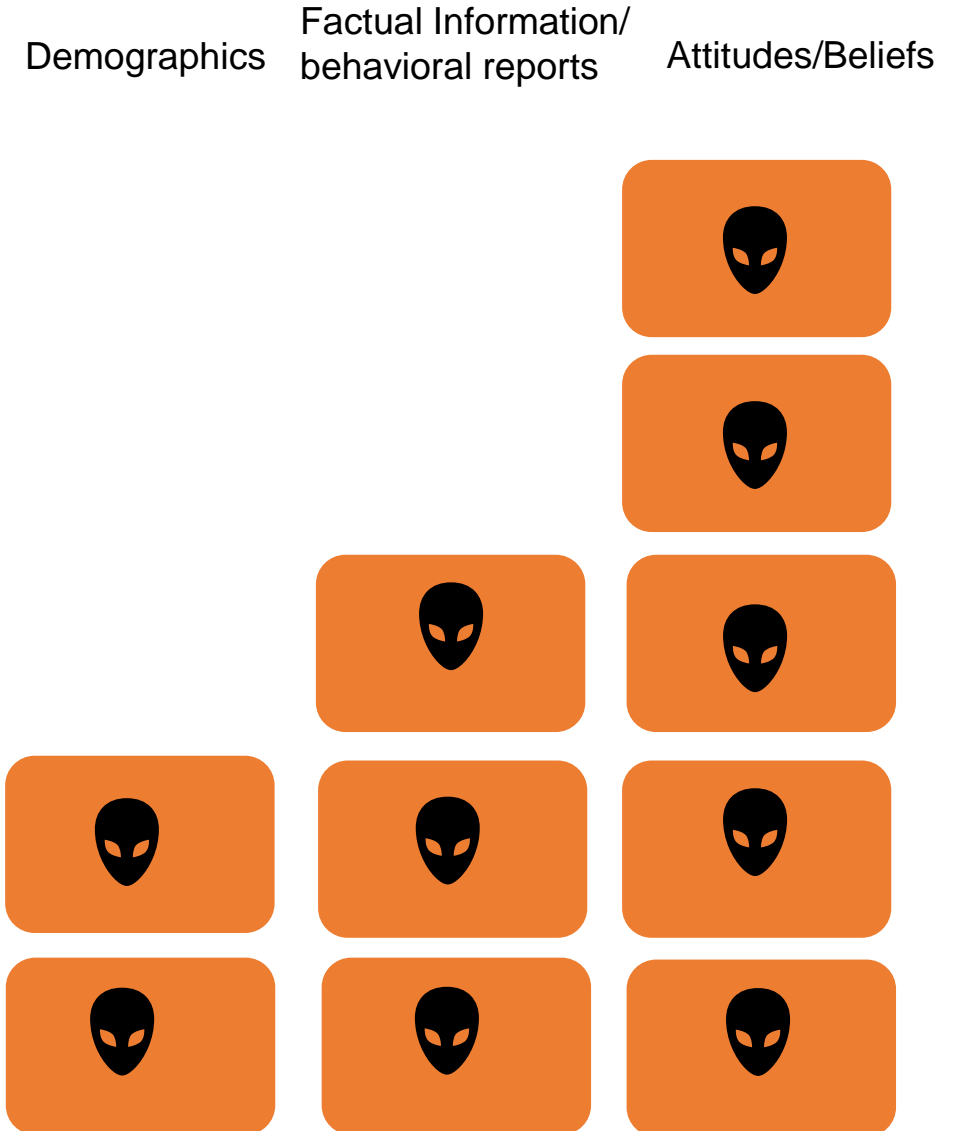
*Do respondents have necessary level of self-awareness?*

*Can respondents remember?*

*Are respondents willing to disclose accurate information?*

*Do all respondents understand the question in the same way?*

J. Bertling, 10/14/2022



# 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, make sure to add a respective 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](mailto:jbertling@ets.org)

[jbertling@fordham.edu](mailto:jbertling@fordham.edu)

[Bertling.Jonas@gmail.com](mailto:Bertling.Jonas@gmail.com)