Rethinking traditional methods of survey validation

Andrew Maul
University of California, Santa Barbara
Talk outline

• Act I: Where we are now

• Act II: How we got here

• Act III: Where we might go next
I: On the shortcomings of traditional approaches to survey validation
Rethinking Traditional Methods of Survey Validation

Andrew Maul

Gevirtz Graduate School of Education, University of California, Santa Barbara

ABSTRACT

It is commonly believed that self-report, survey-based instruments can be used to measure a wide range of psychological attributes, such as self-control, growth mindsets, and grit. Increasingly, such instruments are being used not only for basic research but also for supporting decisions regarding educational policy and accountability. The validity of such instruments is typically investigated using a classic set of methods, including the examination of reliability coefficients, factor or principal components analyses, and correlations between scores on the instrument and other variables. However, these techniques may fall short of providing the kinds of rigorous, potentially falsifying tests of relevant hypotheses commonly expected in scientific research. This point is illustrated via a series of studies in which respondents were presented with survey items deliberately constructed to be uninterpretable, but the application of the aforementioned validation procedures nonetheless returned favorable-appearing results. In part, this disconnect may be traceable to the way in which operationalist modes of thinking in the social sciences have reinforced the perception that attributes do not need to be defined independently of particular sets of testing operations. It is argued that affairs might be improved via greater attention to the manner in which definitions of psychological attributes are articulated and greater openness to treating beliefs about the existence and measurability of psychological attributes as hypotheses rather than assumptions—in other words, as beliefs potentially subject to revision.
Rethinking Traditional Methods of Survey Validation

Andrew Maul

Gevirtz Graduate School of Education, University of California, Santa Barbara

Moving Beyond Traditional Methods of Survey Validation

Andrew Maul

Gevirtz Graduate School of Education, University of California, Santa Barbara

“This is more than the usual wish to transcend one’s predecessors, for it includes a rebellion against the philosophical impulse itself, which is felt as humiliating and unrealistic. It is natural to feel victimized by philosophy, but this particular defensive reaction goes too far. It is like the hatred of childhood and results in a vain effort to grow up too early, before one has gone through the essential formative confusions and exaggerated hopes that have to be experienced on the way to understanding anything. Philosophy is the childhood of the intellect, and a culture that tries to skip it will never grow up.”

—Thomas Nagel (1989)

I argued that traditional methods for the construction and validation of self-report survey-based
Motivation

• Interest in the measurement of “noncognitive” properties has increased in recent years, especially in educational contexts (e.g., ESSA, CORE districts in California)

• Duckworth & Yeager (2015) argue that the enthusiasm is ahead of the science, and that all measures have (different) drawbacks.

• I agree with their analysis, but I think it may not go far enough in diagnosing the most significant barriers to the development of valid measures.
Theories of intelligence

• The *Theory of Intelligence Scale* (e.g., Blackwell, Trzesniewski, & Dweck, 2007) consists of items such as:

“You have a certain amount of intelligence, and you can’t really do much to change it.”

“No matter who you are, you can significantly change your intelligence level.”
The received view

It appears to be commonly believed that survey validation should consist of:

a. estimating reliability (usually via Cronbach’s alpha)

b. (exploratory) factor analysis

c. examination of correlations with other variables

• Issues of intended use and theory-based explanations of behavior are typically not foregrounded.
How far could we push this...?

1. You have a certain amount of intelligence, and you can’t really do much to change it.
1. You have a certain amount of gavagai, and you can’t really do much to change it.

How far could we push this...?
How far could we push this...?

1. Sale mollis qualisque eum id, molestie constituto ei ius.

Strongly disagree  Disagree  Slightly disagree  Slightly agree  Agree  Strongly agree
How far could we push this...?
Methods

• All respondents were Mechanical Turk workers.

• Study 1 ($n = 400$): Theory of Gavagai
  • Study 1b ($n = 100$): “kanin” instead of “gavagai”
  • Study 1c ($n = 100$): “quintessence” instead

• Study 2 ($n = 200$): lorem ipsum

• Study 3 ($n = 200$): blank items
Psychometric properties...

- High reliability
  - $\alpha = .91$ (8 items) for the Theory of Gavagai items
  - $\alpha = .96$ (8 items) for the lorem ipsum items
  - $\alpha = .96$ (8 items) for the blank items

- Good fit to latent variable models
  - EFA suggested a two-factor model for the Theory of Gavagai items, and a one-factor model for the lorem ipsum and blank items
  - All three item sets displayed reasonable fit to a rating scale Rasch model (e.g., most infit meansquares fell between .9 and 1.1)
Psychometric properties...

- Convergent/discriminant validity (?)
  - The Theory of Gavagai item set correlated \((r = .49)\) with the original Theory of Intelligence item set
  - The nonsense item set correlated \((r = -.27)\) with Agreeableness
As an aside:

- In study 2 (lorem ipsum), half the respondents were given the original six-point rating scale, while the other half were given a two-point (agree/disagree) scale
  - $\alpha = .96$ (8 items) for the polytomous items
  - $\alpha = .85$ (8 items) for the dichotomous items
II: The ghosts of operationalism
Post-mortem

• The process of “validating” an instrument is often thought of as separate from the process of defining the property to be measured and articulating hypotheses on the connection between the property and the outcomes of the instrument (Borsboom, 2006)

• The legacy of operationalism (Bridgman, 1927; Stevens, 1946) has reinforced the perception that as long as one has an instrument, one has a property (and a measure) by fiat

• Greater attention to definitional issues could help provide theoretical justification for choices made “downstream” (e.g., item development, psychometric modeling)
Defining a measurable property

Self-control is “defined as the regulation of attention, emotion, and behavior when enduringly valued goals conflict with more immediately pleasurable temptations” (Duckworth & Yeager, 2015, p.239)

- This is not transparently a definition of an property of persons, and is therefore not transparently a definition of anything measurable about persons.
Disposition, tendency, capacity, ability...?

• Proposal #1: self-control is the *ability* to regulate attention, emotion, and behavior...

• Proposal #2: self-control is the *tendency* to regulate attention, emotion, and behavior...

• Cronbach (1949) differentiated between assessments of *maximal* and *typical* behavior.

• Related: to what extent is self-control ontologically subjective ("the perception is the reality")?
  • “I am a hard worker”
  • “I worked __ hours yesterday”
Relating a property to observations

• Hypothesized cause of observations or inductive summary of observations? (e.g., Edwards & Bagozzi, 2000)

• If the former, does the property cause only between-person variation in observations, or also within-person variation? (e.g., Borsboom, Mellenbergh, & van Heerden, 2003)
Relating a property to observations

Is there a causal claim?
- Yes
  - Is the claim at the within person level?
    - Yes
      - Is the theory at the process level?
        - Yes
          - Different causal processes?
            - Yes
              - Multiple process causation within persons.
            - No
              - Single process causation within persons.
        - No
          - Black box causation within persons.
    - No
      - Black box causation between persons.
- No
  - Explicit no causation

Functional attributes

Mechanistic attributes

Non-committal take on causation

Latent Vars. Models w/Cross-sectional data

Latent Vars. Models w/Time series data

Evidence based on response processes

Explanatory IRMs CDMs
III: Where do we go from here?
What is (quantitative) measurement, and what is it for?

• Measurement:
  • involves formal representation of empirical relations via numerical relations, where
  • the results are expected to *carry meaning*, and are
  • valued because they can be *acted upon*.

• In other words, measurement involves *syntax, semantics, and pragmatics*

• From this, the criteria for validity are straightforward: the information obtained must be of sufficiently high quality to support the intended use of the instrument.
Where do we go from here?

• Suppose we wish to increase our knowledge about some property of human beings. With which of the following statements do you most strongly agree?
What are we willing to give up?

a) We must measure the property using a self-report survey-based instrument with Likert-type response options.

b) We must measure the property using a self-report survey-based instrument, but the response options need not follow a Likert-like format.

c) We believe that measurement of the property is both possible and desirable, but we need not use a survey.

d) We are not sure if measurement is (presently) possible, feasible, desirable, and/or optimal.
What are we willing to give up?

a) We must measure the property using a self-report survey-based instrument with Likert-type response options.

- Options include:
  - Following best practices in designing and qualitatively evaluating items (e.g., Gehlbach, & Brinkworth, 2011)
  - Anchoring vignettes (e.g., Bolt & Kim, 2014)
What are we willing to give up?

b) We must measure the property using a self-report survey-based instrument, but the response options need not follow a Likert-like format.

• Options include:
  • Ipsative or other forced-choice response options (e.g., Kyllonen, 2015)
  • scaling methods developed by Thurstone, Coombs, Guttman, and others (e.g., McGrane & Nowland, 2018; for a modern perspective on Guttman items, see also Wilson, 2005)
What are we willing to give up?

c) We believe that measurement of the property is both possible and desirable, but we need not use a survey.

- Options now include performance tests, structured interviews, scoring of essays or other specific artifacts, or behavioral observations...
What are we willing to give up?

a) We must measure the property using a self-report survey-based instrument with Likert-type response options.

b) We must measure the property using a self-report survey-based instrument, but the response options need not follow a Likert-like format.

c) We believe that measurement of the property is both possible and desirable, but we need not use a survey.

d) We are not sure if measurement is (presently) possible, feasible, desirable, and/or optimal.
What are we willing to give up?

“Although reason and imagination also advance knowledge... only measurement makes it possible to observe patterns and... to put one’s guesses about what is and is not true to the test” (Duckworth & Yeager, 2015, p. 237, emphasis added)

• This is simply false.
  • It ignores the value of qualitative approaches to inquiry (e.g., Green and Bloome, 2004; Gumperz and Cook-Gumperz, 2007; Street, 2014)
  • It’s not even true in the physical sciences (see, e.g., McGrane, 2015; Sherry, 2011)
  • Whether any given act of inquiry requires (or can even benefit from) measurement depends on the nature of the phenomena (e.g., Michell, 1999, Humphry, 2013, Kyngdon, 2013, Markus & Borsboom, 2013)
"It is natural to feel victimized by philosophy, but this particular defensive reaction goes too far. It is like the hatred of childhood and results in a vain effort to grow up too early, before one has gone through the essential formative confusions and exaggerated hopes that have to be experienced on the way to understanding anything. Philosophy is the childhood of the intellect, and a culture that tries to skip it will never grow up."

- Thomas Nagel
Thank you for your time!

amaul@ucsb.edu
References


References


References


