Construct Validation in Social and Personality Research: Current Practice and Recommendations (Flake et al., 2017)

ReproducibiliTea Week 3
15.04.2021
Shermain

How would you measure these?

Observed/ Non-latent variables

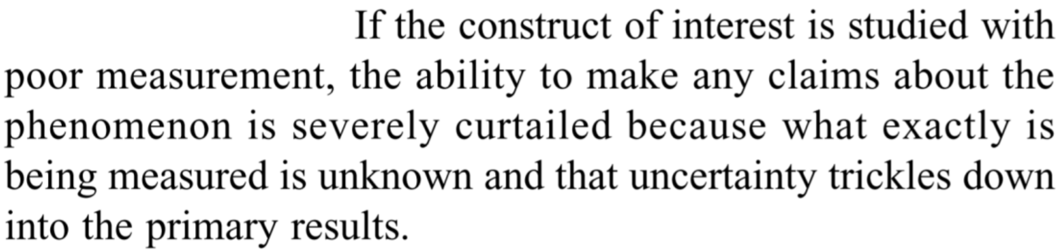
- Temperature
- Mobile phone usage
- Income

Non-observed/Latent variables

- Depression
- Happiness
- Motivation

Theory → Measure (e.g. likert scales) → **Construct validation**

Why is construct validation important?



If the construct of interest is studied with poor measurement, the ability to make any claims about the phenomenon is severely curtailed because what exactly is being measured is unknown and that uncertainty trickles down into the primary results.

Best Practices

Substantive

Lit review, content relevance (e.g. cognitive interview), item development

Structural

Item analysis, factor analysis, reliability, measurement invariance

External

Convergent and discriminant, predictive/criterion

Which of these have you done/ typically see being done?

Best Practices

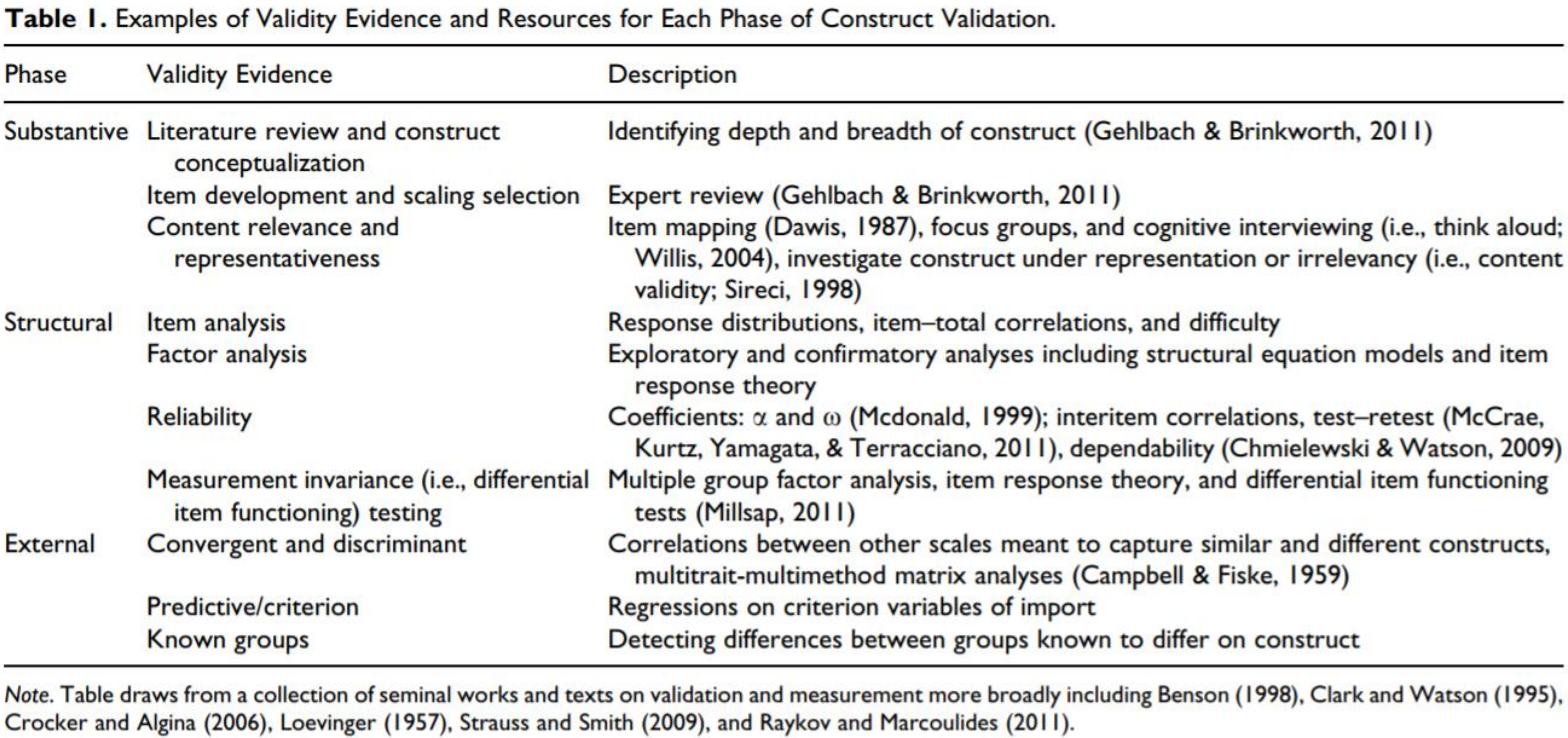
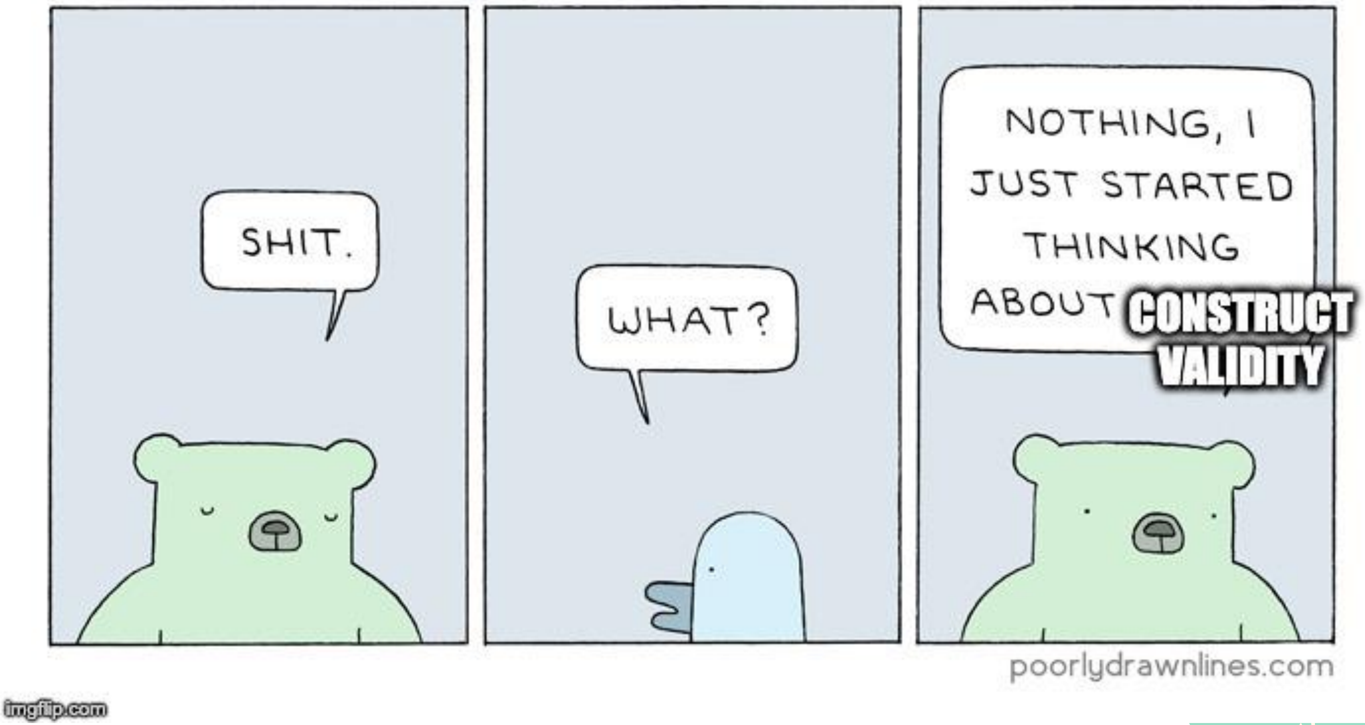


Table 1. Examples of Validity Evidence and Resources for Each Phase of Construct Validation.

Phase	Validity Evidence	Description
Substantive	Literature review and construct conceptualization	Identifying depth and breadth of construct (Gehlbach & Brinkworth, 2011)
	Item development and scaling selection	Expert review (Gehlbach & Brinkworth, 2011)
	Content relevance and representativeness	Item mapping (Dawis, 1987), focus groups, and cognitive interviewing (i.e., think aloud; Willis, 2004), investigate construct under representation or irrelevancy (i.e., content validity; Sireci, 1998)
Structural	Item analysis	Response distributions, item–total correlations, and difficulty
	Factor analysis	Exploratory and confirmatory analyses including structural equation models and item response theory
	Reliability	Coefficients: α and ω (Mcdonald, 1999); interitem correlations, test–retest (McCrae, Kurtz, Yamagata, & Terracciano, 2011), dependability (Chmielewski & Watson, 2009)
	Measurement invariance (i.e., differential item functioning) testing	Multiple group factor analysis, item response theory, and differential item functioning tests (Millsap, 2011)
External	Convergent and discriminant	Correlations between other scales meant to capture similar and different constructs, multitrait-multimethod matrix analyses (Campbell & Fiske, 1959)
	Predictive/criterion	Regressions on criterion variables of import
	Known groups	Detecting differences between groups known to differ on construct

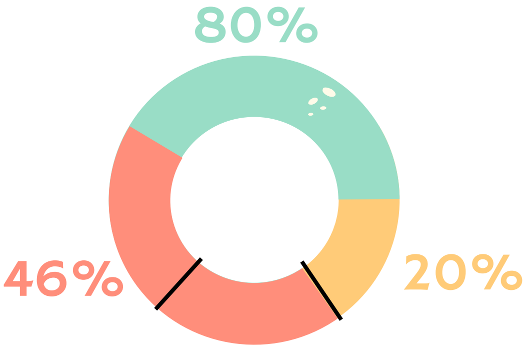
Note. Table draws from a collection of seminal works and texts on validation and measurement more broadly including Benson (1998), Clark and Watson (1995), Crocker and Algina (2006), Loevinger (1957), Strauss and Smith (2009), and Raykov and Marcoulides (2011).

Current Practice/Problems



Current Practice/Problems

- More than 80% of social and personality psychology research include latent variable measurement
- Almost half do not reference previous validation (appear developed on the fly; new)
- Half of these only report Cronbach's α
- Valid measurement is a **necessary prerequisite** to the interpretation of results
- Evidence is required to reflect accuracy of measure of purported construct of interest



Current Practice/Problems

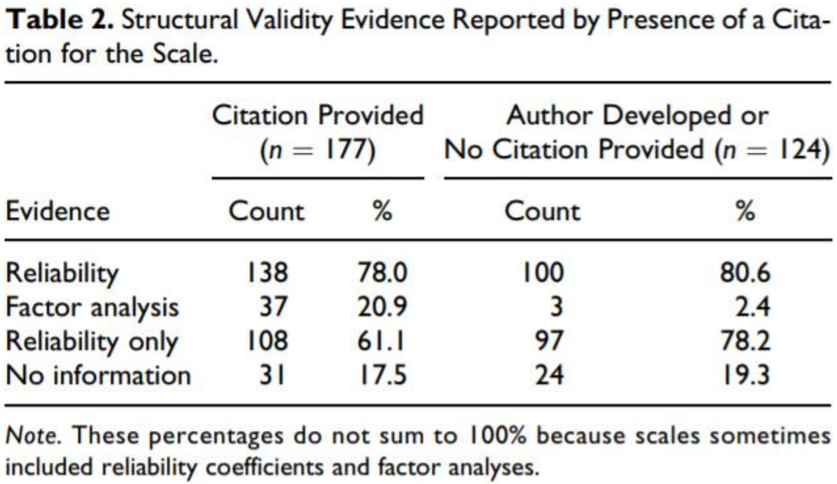


Table 2. Structural Validity Evidence Reported by Presence of a Citation for the Scale.

	Citation Provided (n = 177)		Author Developed or No Citation Provided (n = 124)	
Evidence	Count	%	Count	%
Reliability	138	78.0	100	80.6
Factor analysis	37	20.9	3	2.4
Reliability only	108	61.1	97	78.2
No information	31	17.5	24	19.3

Note. These percentages do not sum to 100% because scales sometimes included reliability coefficients and factor analyses.

- Structural validity evidence for scales with 2 or more items
- Cronbach's α was the most common reliability coefficient provided, comprising 73% (n = 222) of reported reliability information, with interitem correlations representing 4%, the remaining scales did not report reliability information
- However, there is a gross misuse of Cronbach's α in the literature

Current Practice/Problems

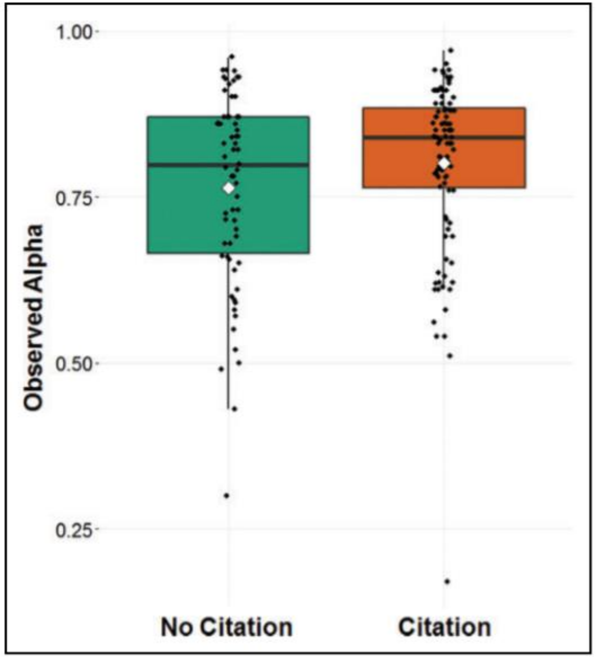


Figure 1. Boxplots of the α distributions for both novel and previously developed scales.

- The distribution of Cronbach's α for whether the scale has a citation provided
- Smaller variability in reliability for cited scales

- "Many constructs studied lack appropriate validation, which will contribute to questionable conclusions and difficulty of subsequent research to replicate."

Current Practice/Problems

- Big theories, small scales (**Poor construct representativeness**)
- 30% of scales have 1 item, and most developed scales have less than 3 items
- Example:

  Status – a multidimensional construct consisting of wealth, social affiliation, and prestige

  This will be difficult to capture with a short 2-3 item scale representing status

- However, sometimes you *need/want* a short scale
- **The problem is not short scales, but the lack of validation in these scales**
- Include multiple sources of validity evidence (content, convergent, predictive), replication, use case scenario (Beymer et al., 2021)

Limitations/Misuse of Cronbach's α

Assumptions

Single factor, equal factor loadings.
McDonald's omega should be reported in the case of unequal factor loadings.

Unidimensionality

Misinterpretation of α as a measure of unidimensionality. Authors combine scales and only report α.

Over-reliance

Criterion for scale use; Item selection; Justify item removal. Should not be used in expense of other CV evidence.

Recommendations

1

Measurement

Measurement properties should be valid before interpretation of results

2

Ongoing Validation

Always validate your scales even if it is an existing scale (but used with a different population)

3

Content

Ensure construct representation and relevance. Broad constructs will generally require longer scales.

4

Cronbach's α

Halt the sole and incorrect use of coefficient α

Formal training on CTT/measurement!

Your thoughts

01 What are some bad & good measurement practices have you done/seen done?

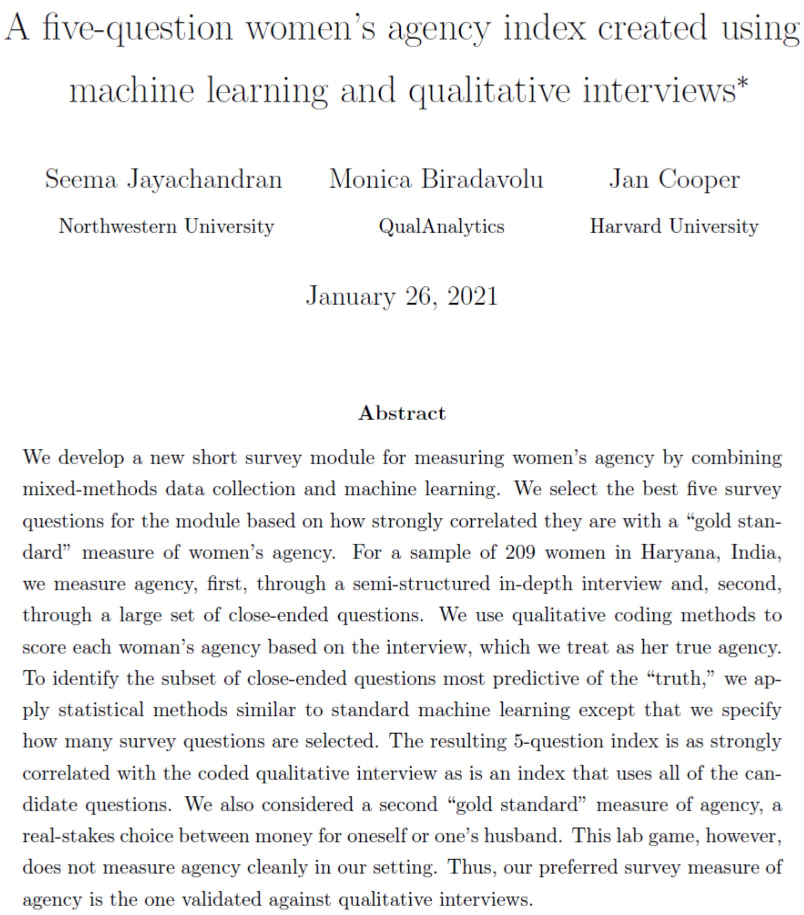
02 What are some other problems/limitations you've faced with scale development?

03 Other than education/training in measurement, what else could be done?

04 General thoughts?

- Jayachandran et al. (2021) proposed using qualitative interview methods to generate "gold standards" measure of their construct of interest

- "The approach is to conduct semi-structured interviews about a complex construct (e.g., women's agency), code them, and use machine learning to choose the (say) 5 survey questions, from among a large set of contenders, that best predict the "gold standard" measure."
  (https://twitter.com/seema_econ/status/1355204891275268108?s=20)



A five-question women's agency index created using machine learning and qualitative interviews*

Seema Jayachandran — Northwestern University
Monica Biradavolu — QualAnalytics
Jan Cooper — Harvard University

January 26, 2021

Abstract

We develop a new short survey module for measuring women's agency by combining mixed-methods data collection and machine learning. We select the best five survey questions for the module based on how strongly correlated they are with a "gold standard" measure of women's agency. For a sample of 209 women in Haryana, India, we measure agency, first, through a semi-structured in-depth interview and, second, through a large set of close-ended questions. We use qualitative coding methods to score each woman's agency based on the interview, which we treat as her true agency. To identify the subset of close-ended questions most predictive of the "truth," we apply statistical methods similar to standard machine learning except that we specify how many survey questions are selected. The resulting 5-question index is as strongly correlated with the coded qualitative interview as is an index that uses all of the candidate questions. We also considered a second "gold standard" measure of agency, a real-stakes choice between money for oneself or one's husband. This lab game, however, does not measure agency cleanly in our setting. Thus, our preferred survey measure of agency is the one validated against qualitative interviews.