MONASH
EDUCATION

Almost everything you need to know about evaluating and developing scales for use in special populations

Shane Costello PhD MAPS

Shane Costello

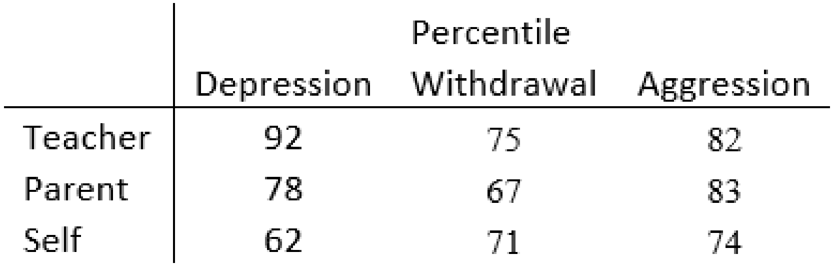
- Bachelor of Science (Psychology)
- Postgraduate Diploma of Psychology
- Master of Psychology (Educational and Developmental)
- Doctor of Philosophy
- Worked in schools (K-12) since 2013
- Monash University lecturer and researcher since 2014
  - Master of Educational and Developmental Psychology
  - Master of Counselling
  - Master of Professional Psychology (Deputy Course Leader)
- Previously worked in disability, aged and community care
- Previous scale projects in inclusive education, clinical education, stages of change, cognitive abilities, cognitive style, empathy, trauma, personality, occupational interests

Objectives

- Describe the importance of using robust measurement tools in both research and clinical applications
- Describe a framework for developing, evaluating, and validating measurement tools
- Demonstrate the process of evaluating and improving a measurement tool
- Demonstrate the process of developing a new assessment tool for a special population

Importance of measurement – clinical practice

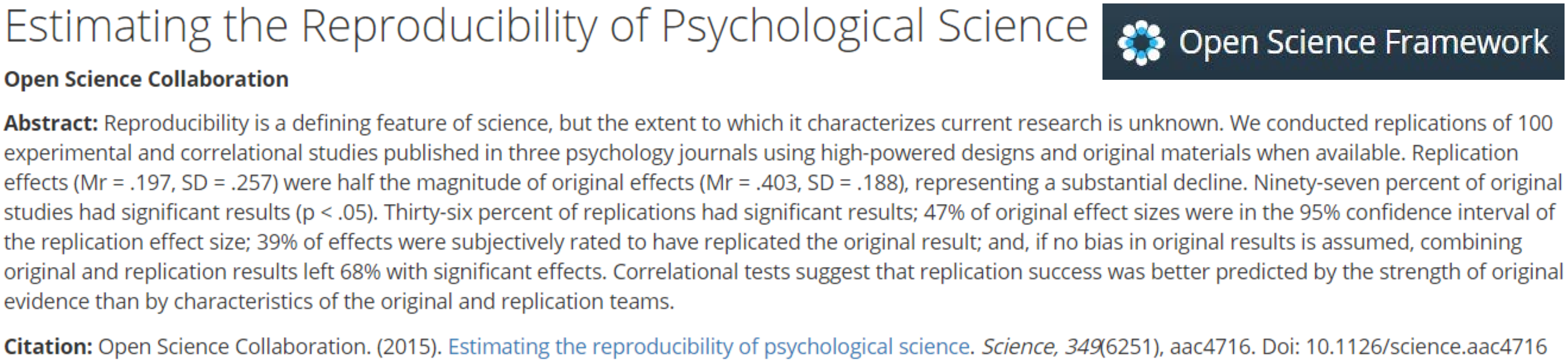
- The case of Stephen
  - Nine year old boy referred to university clinic because teacher was concerned about low mood. Presented quite flat and unenthusiastic during cognitive testing



		Percentile	
	Depression	Withdrawal	Aggression
Teacher	92	75	82
Parent	78	67	83
Self	62	71	74

- Three different measures across three raters – assuming all rated honestly, which are accurate?
- What are the possible consequences of accepting the wrong measure?

Importance of measurement – research



Estimating the Reproducibility of Psychological Science

Open Science Collaboration

Abstract: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects (Mr = .197, SD = .257) were half the magnitude of original effects (Mr = .403, SD = .188), representing a substantial decline. Ninety-seven percent of original studies had significant results (p < .05). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Citation: Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716

- Better measures have smaller standard errors, which reduces sampling effects and makes it less likely for you to conclude a significant difference when there is none
- What are some other consequences of poor measurement tools in research?

The biggest problem in measurement…..

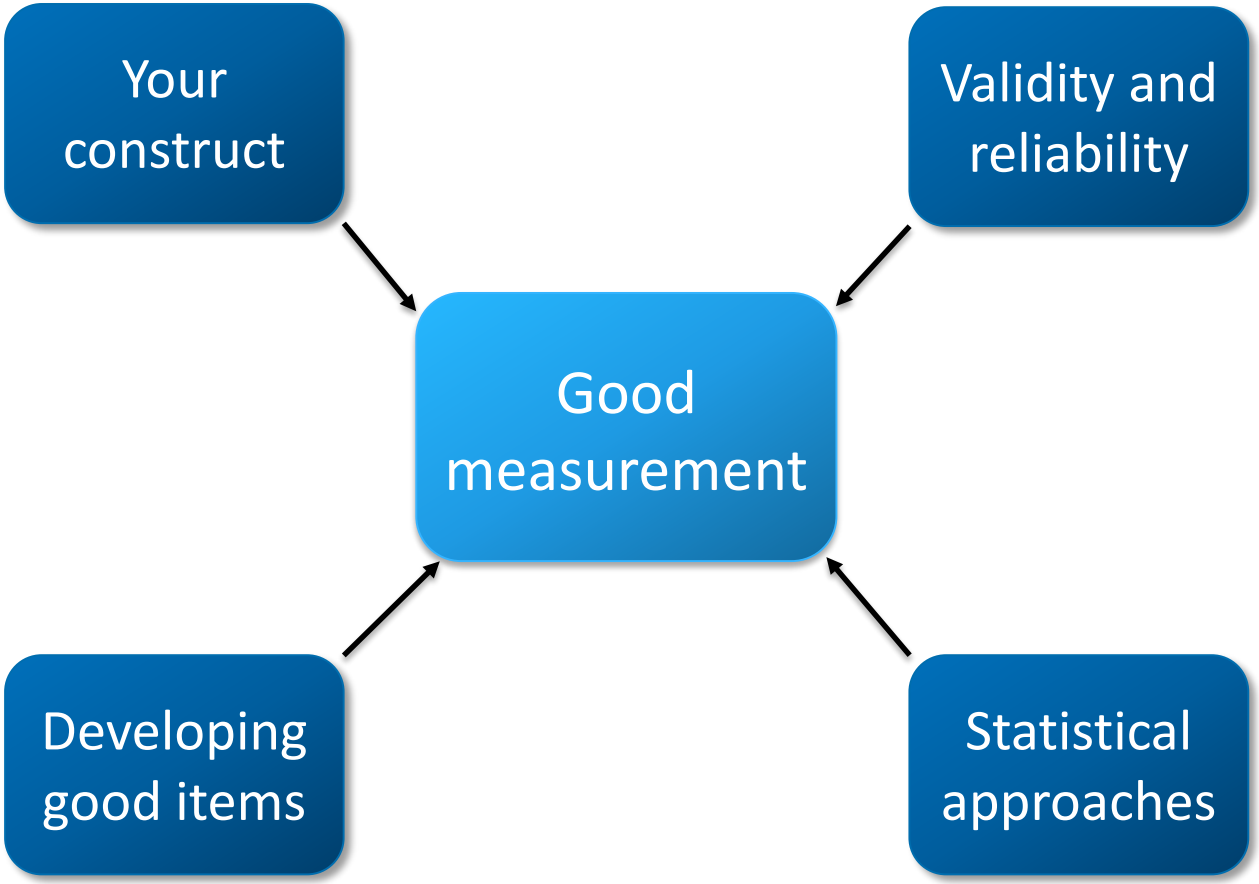
"The biggest problem in the design, evaluation and validation of measurement tools is a lack of theory. Ignoring the theoretical frameworks of the constructs themselves; limited application of the theory of validity and reliability; being unaware of the theories of item development; and not understanding statistical theories and methodologies."

Shane Costello, Michigan State University, Jan 2017

What about special populations?

- Every population is special (ie unique, differing in some way from the general population)
  - Ignoring that uniqueness leads to error
  - More error → poor measurement
  - Poor measurement → bad research, bad outcomes for clients
- Good measures are developed specifically for special populations, or adapted carefully to suit.
- Good measures are developed in collaboration with the special population, not just defined by “experts”

The meta-theoretical framework of scale development



Your construct

Validity and reliability

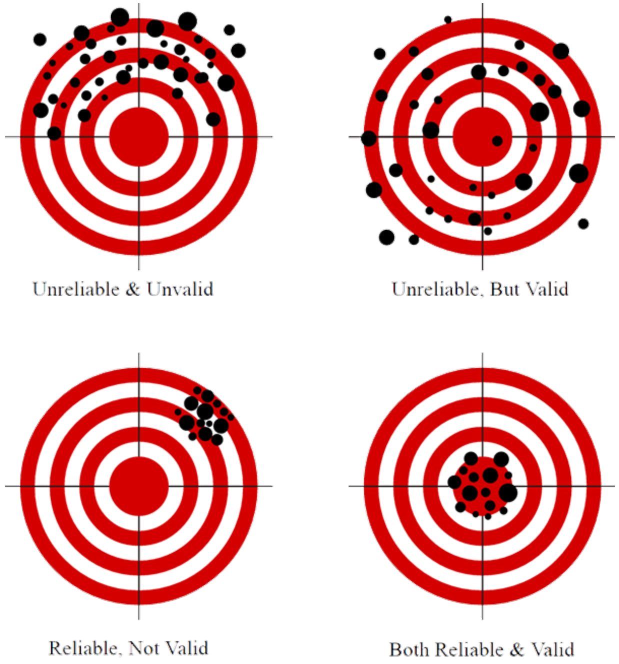
Good measurement

Developing good items

Statistical approaches

Validity and reliability

- *Validity* – does the scale measure what it purports to measure?
- *Reliability* – does the scale consistently measure the same thing?



Unreliable & Unvalid

Unreliable, But Valid

Reliable, Not Valid

Both Reliable & Valid

Validity and reliability

Messick's (1995) model of construct validity

- *Content relevance and representativeness*
  - Determining the boundaries of the construct, and ensuring that the breadth of the construct is represented in the tool
- *Substantive theories, process models, and process engagement*
  - Ensuring that the tool is grounded in theory, and draws on processes as well as content (eg a maths question that tests ability to do maths, not just memorization)
- *Scoring models as reflective of task and domain structure*
  - Scoring should be consistent with what is known about the construct
- *Generalizability and the boundaries of score meaning*
  - Does the test generalize? Upper/lower limits of measurement?
- *Convergent and discriminant validity*
  - Is the test related to what it should be, and not related to what it shouldn't be?
- *Consequences as validity*
  - What are the intended consequences of the use of the test? Are there any unintended or negative consequences that may occur?

Construct underrepresentation – failing to include everything needed in a measure

Construct-irrelevant variance – measuring things you didn't intent to

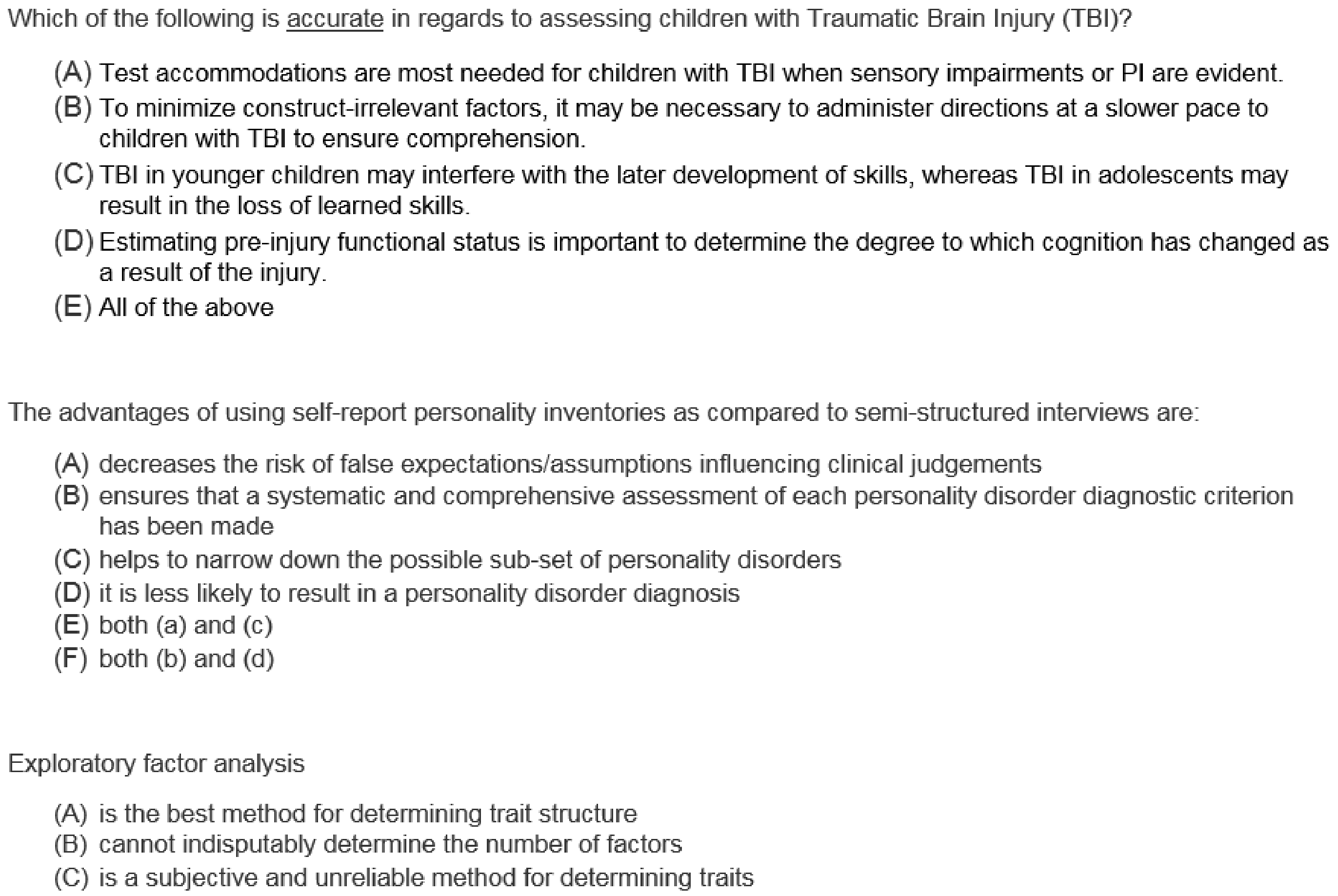
Developing good items

Multiple choice performance measures

- Have three components
  - Stem, or question statement. Ideally contains all of the information needed for a participant to know the whole question. Can include “red herrings” or additional unused information (but be careful – what are you trying to test?)
  - The correct answer. Self-explanatory
  - Distractor answers. Ideally these are all plausible answers
- Some important points
  - Long questions with short answers are always better
  - Don’t give away answers (obviously incorrect choices)
  - Avoid “all of the above”, “none of the above”, and “both (a) and (b)”
  - Consistency in number of answer choices is usually more reliable (typically A, B, C, or D)
  - If using True/False, many more questions are needed
  - Absolute responses (never/always) are less likely to be correct, and therefore less plausible
  - Aim for a spread of questions with a correct range of 25-75%

Developing good items

What is wrong with these items?



Which of the following is accurate in regards to assessing children with Traumatic Brain Injury (TBI)?

(A) Test accommodations are most needed for children with TBI when sensory impairments or PI are evident.
(B) To minimize construct-irrelevant factors, it may be necessary to administer directions at a slower pace to children with TBI to ensure comprehension.
(C) TBI in younger children may interfere with the later development of skills, whereas TBI in adolescents may result in the loss of learned skills.
(D) Estimating pre-injury functional status is important to determine the degree to which cognition has changed as a result of the injury.
(E) All of the above

The advantages of using self-report personality inventories as compared to semi-structured interviews are:

(A) decreases the risk of false expectations/assumptions influencing clinical judgements
(B) ensures that a systematic and comprehensive assessment of each personality disorder diagnostic criterion has been made
(C) helps to narrow down the possible sub-set of personality disorders
(D) it is less likely to result in a personality disorder diagnosis
(E) both (a) and (c)
(F) both (b) and (d)

Exploratory factor analysis

(A) is the best method for determining trait structure
(B) cannot indisputably determine the number of factors
(C) is a subjective and unreliable method for determining traits

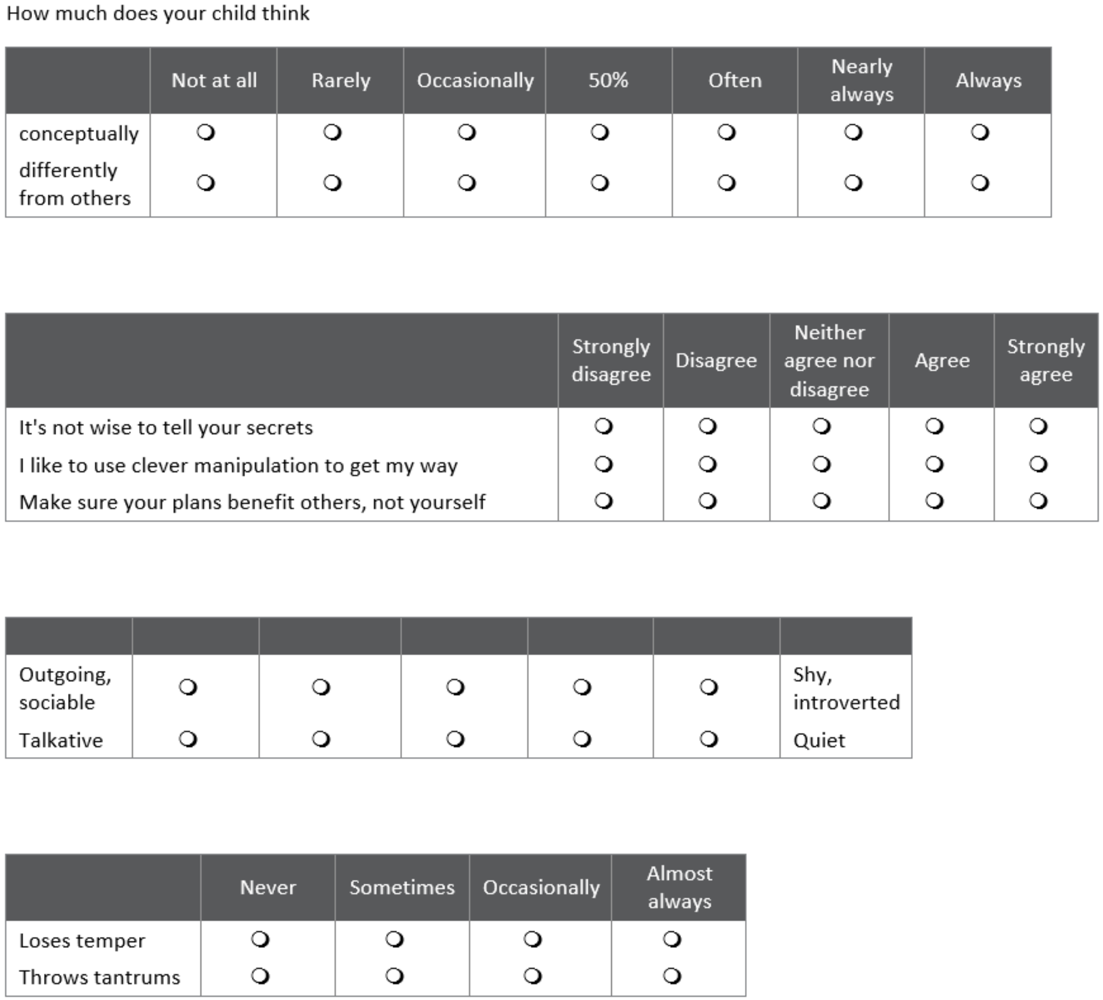
Developing good items

Preference measures

- Have two components
  - Stem, or question statement
  - Response scale
- Some important points
  - Question statements should be clear, and only contain one statement
  - Is the scale unipolar or bipolar? Link to construct theory
  - Number of response categories range from 2 to ∞
  - Ability to distinguish between categories increases with age and education
  - General rule: children (3), adolescents (5), adults (7), higher educated (9)
  - Even or odd number of response categories (is there a meaningful midpoint?)
  - Avoid negative questions, but meaningfully opposite items are fine (shy vs outgoing)

Developing good items

Preference measures



How much does your child think

	Not at all	Rarely	Occasionally	50%	Often	Nearly always	Always
conceptually	○	○	○	○	○	○	○
differently from others	○	○	○	○	○	○	○

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
It's not wise to tell your secrets	○	○	○	○	○
I like to use clever manipulation to get my way	○	○	○	○	○
Make sure your plans benefit others, not yourself	○	○	○	○	○

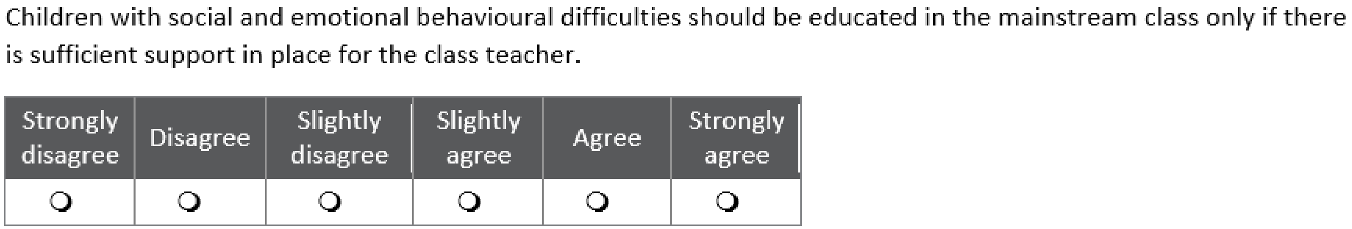
Outgoing, sociable	○	○	○	○	○	Shy, introverted
Talkative	○	○	○	○	○	Quiet

	Never	Sometimes	Occasionally	Almost always
Loses temper	○	○	○	○
Throws tantrums	○	○	○	○

MONASH University

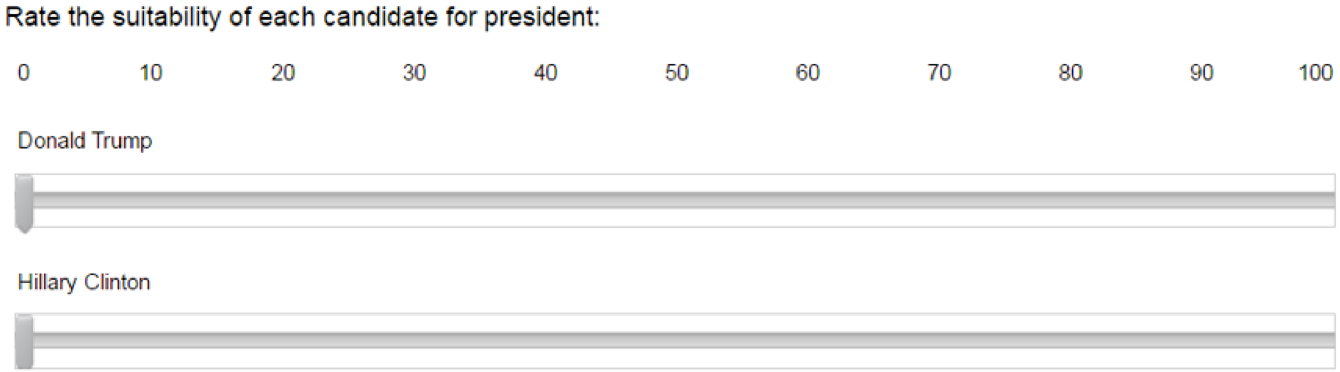
Developing good items

What is wrong with these items?



Children with social and emotional behavioural difficulties should be educated in the mainstream class only if there is sufficient support in place for the class teacher.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree
○	○	○	○	○	○

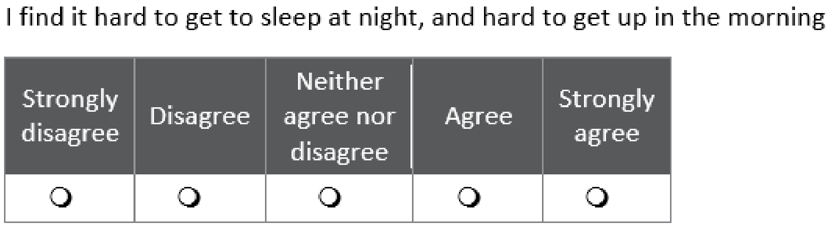


Rate the suitability of each candidate for president:

0 10 20 30 40 50 60 70 80 90 100

Donald Trump

Hillary Clinton



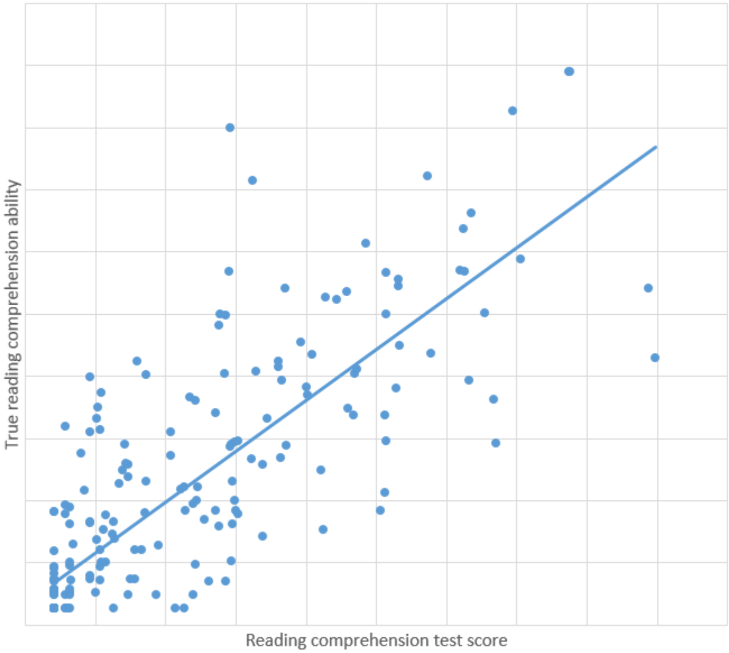
I find it hard to get to sleep at night, and hard to get up in the morning

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
○	○	○	○	○

Statistical approaches

Classical Test Theory

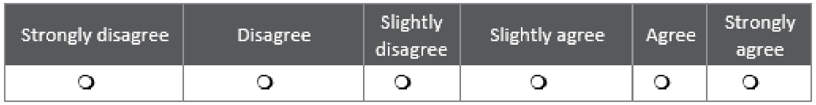
- Observed score = true score + error
  - Assumes a linear relationship between measurement and construct
  - Assumes linear progression through response categories
  - Used for factor analysis (exploratory and confirmatory)
- Some important points
  - Items should be of similar difficulty (performance) or intensity (preference)
  - Cronbach's alpha increases with number of items, so be aware that a long scale may not be as good as it looks
  - Specific scale length is 4 to 6 items (but much longer is common and often needed for reliability)
  - Can easily test multidimensional models (eg validity of separate subscales)
  - Does not work well with dichotomous items
  - Affected by sampling bias



Statistical approaches

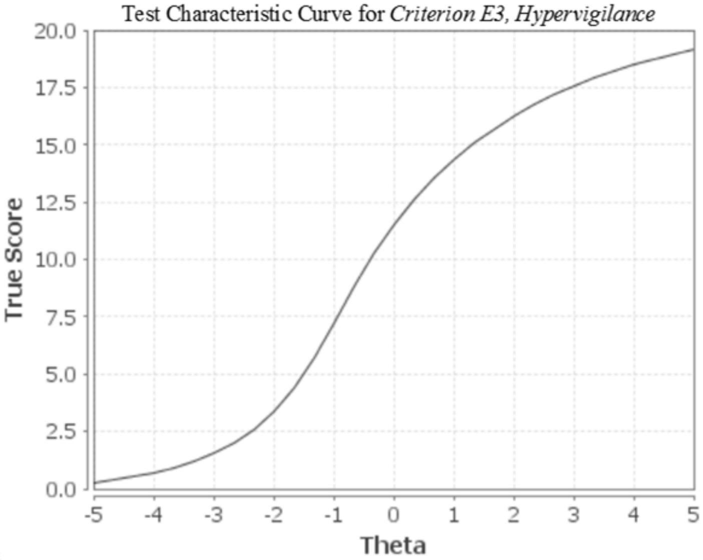
Item Response Theory

- Probability of correct answer = $\dfrac{\text{exp (ability – item difficulty)}}{\text{1 + exp (ability – item difficulty)}}$
  - Because measurement is based on probability, no linear relationship is assumed
  - Allows for differences in distance



Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree
○	○	○	○	○	○

- Some important points
  - Items should range in difficulty from easy to hard (or least likely to most likely to be endorsed)
  - Scale should be unidimensional
  - “Reliability” increases with more discriminating items, which reduces the standard error of measurement
  - Longer scales aren’t necessarily better (or more reliable)
  - Can model difficulty, discrimination, guessing, and carelessness to improve measurement precision
  - Less affected by sampling bias



Developing a measurement tool – the process

The development phase

NEED

Identify the need for a new test or measure. Review literature for existing measures and note problems. Decide on theoretical framework of construct.

SPECS

Define the ideal "end product". Is it brief/comprehensive? For test/retest, single use, different populations. Statistical framework etc

INITIAL

Develop an initial pool of items. Use expert consensus. Target population evaluate for ecological validity (focus groups, cognitive interviewing)

Developing a measurement tool – the process

The testing phase

SMALL	Conduct a pilot study using the initial items. This study should have 50-100 participants.
EVAL	Preliminary analysis using chosen statistical framework. Consider poor performing items – why? Interviews/focus groups for improvements.
LARGE	Conduct a large study, aiming to sample at least 300 participants.
FINAL	Final analysis using chosen statistical framework.

Developing a measurement tool – the process

The consolidation phase

SCALE

Convert raw scores to scaled scores using chosen statistical framework.

NORM

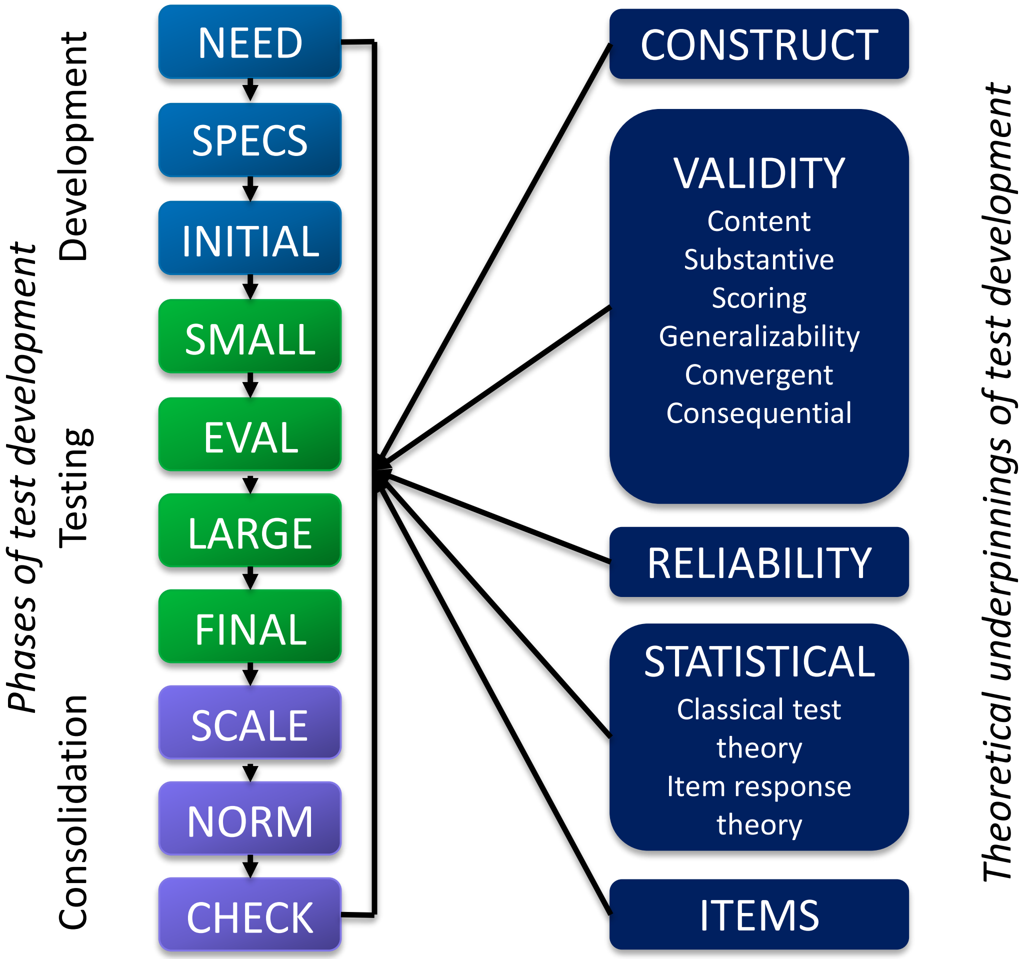
Conduct norming study if required. Sampling needs to account for population characteristics and prevalence of sub-populations

CHECK

Conduct final checks of reliability/test-retest. Convergence/discrimination with external measures

Adapted from Roodenburg (2006) and McGrew (2009)

Developing a measurement tool – the complete framework



Phases of test development

Development

NEED

SPECS

INITIAL

Testing

SMALL

EVAL

LARGE

FINAL

Consolidation

SCALE

NORM

CHECK

CONSTRUCT

VALIDITY
Content
Substantive
Scoring
Generalizability
Convergent
Consequential

RELIABILITY

STATISTICAL
Classical test theory
Item response theory

ITEMS

Theoretical underpinnings of test development

Further reading

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319. doi: 10.1037/1040-3590.7.3.309

De Raad, B., & Hendriks, A. A. J. (1997). A psycholexical route to content coverage in personality assessment. *European Journal of Psychological Assessment*, 13(2), 85-98. doi: 10.1027/1015-5759.13.2.85

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4 ed.). Thousand Oaks, CA: SAGE Publications.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286-299. doi: 10.1037/1040-3590.7.3.286

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks CA: SAGE Publications.

Hogan, N., Costello, S., Boyle, M., & Williams, B. (2015). Measuring workplace trauma response in Australian paramedics: an investigation into the psychometric properties of the Impact of Event Scale. *Psychology Research and Behavior Management*, 8, 287-294. doi: 10.2147/PRBM.S96647

Jacobs, K. E., & Costello, S. (2013). An Initial Investigation of an Australian Adaptation of the Multidimensional Aptitude Battery — II. *The Educational and Developmental Psychologist*, 30(1), 84-102. doi: 10.1017/edp.2013.9

Messick, S. (1995). Validity of psychological assessment; Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741/749.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46(1), 1-18. doi: 10.1348/014466506x96931

Roodenburg, J. (2006). *Personality-centred cognitive style: Construct modelling based on an exploration of teachers' perceptions of individual differences in student thinking.* (Doctor of Philosophy), University of Melbourne, Melbourne Australia.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika*, 74(1), 107-120. doi: 10.1007/s11336-008-9101-0

Watt, D., Hopkinson, L., Costello, S., & Roodenburg, J. (2016). Initial validation and refinement of the Hierarchical Inventory of Personality for Children in the Australian context. *Australian Psychologist*. doi: 10.1111/ap.12213

Slides available from

shanecostello.net

Questions?