A. Alexander Beaujean

Associate Professor of Psychology & Neuroscience

Unmuddled Measurement

Straight talk about psychological measurement and score interpretation

Measurement

Motivation

Definition (Measurement)

Measurement is a both conceptual and experimental process implementing a morphic property value assignment able to produce information on a predefined property with a specified and provable level of objectivity and inter-subjectivity. (Mari, Carbone, & Petri, 2015, p. 219)

Motivation

Upshot

Measurement is supposed to represent a relation that exists in the "real world"

Motivation

Definition (Quantity)

A quantitative phenomenon is one where there the relations among distinguishable elements have both *order* and *additivity*.

Motivation

Before representing attributes we need empirical (experimental) investigations

Motivation

Psychological attributes

- Any empirical reason to believe they are quantitative?
- Or are these just qualitative distinctions?

Motivation

Can use numbers to represent things

- **Does not** make numbers have properties we are accustomed to them having

Motivation

Upshot

1. Understand (empirical) nature of the attribute, **then**
2. Represent attribute via measurement

Measurement

Measurement is a *process*

- *Both* conceptual and experimental
- Process outcome: **measurement result**

Measurement

Measurement Goal

Goal of measurement: provide *information* about a specific *attribute* that describes the empirical world effectively.

Concepts

Definition (Concept)

Concepts are elements of language used to denote and organise phenomena (Maraun & Gabriel, 2013).

Concepts

Concepts

Concepts are human creations/inventions.

Concepts

Concept meaning

- Fixed by linguistic rules
  - Determine its range of correct employments
- Meaning manifested: range of correct employments

Concepts

Measurement concepts of interest: object attributes

Concepts

Clarify concepts and their constitutive referents **before** beginning measurement

Empirical Component

Attributes: not directly measurable (they are concepts)

- Instead, measure attribute *manifestations*

Empirical Component

Attribute manifestations: divide into classes that are **mutually exclusive** and **exhaustive**

- Classes
  - Do **not** overlap
  - Contain all possible manifestation of attribute

Empirical Component

Definition (Equivalence Classes)

Equivalence classes are things that are alike (i.e., relations among them are indistinguishable) regarding particular phenomena of interest

Empirical Component

Measurement: represent what is known about relations among distinguishable attribute manifestations

Empirical Component

Empirical Information in Measurement

Empirical information has to be known about attribute *before* representing it

Empirical Component

Upshot

Empirical vs. Conceptual

- Conceptual aspect of measurement: Involves no discovery
  - Rule based
- Empirical aspect of measurement: Involves discovery
  - Discover things about attribute (manifestations)

Empirical Component

Discover information about attribute manifestation relations $\Rightarrow$ represent information symbolically (e.g., numerals)

Empirical Component

Measurement: create map from real world (i.e., empirical relations) to model (i.e., numeral relations)

Empirical Component

Measurement result: set of values

Empirical Component

Measurement Values

Measurement values: typically represented as random variable

- Variables **not** the same as phenomena they represent

Empirical Component

Attribute manifestations need to meet certain conditions to be measurable

- Representation
- Uniqueness
- Meaningfulness
- Scaling
- (Uncertainty)

Representation

What do we know about the equivalence classes?

- Elements in different classes

- Are distinction aspects of phenomenon?
- Represent more or less of the phenomenon? (ordinality)
- Can be combined to represent a distinct class? (additivity)

Uniqueness

Definition (Measurement Scale)

A measurement scale is the particular way we assign numbers to measure the attribute.

Uniqueness

Typically ≥ 1 set of symbols we can use

- More empirically-discovered relations to preserved $\Rightarrow$ fewer unique symbol systems available

Uniqueness

Uniqueness: closely related to value transformation

- Any symbol transformations we make must keep the representation intact (i.e., admissible transformation)

Uniqueness

Scale Type	Description
Absolute	Counts
Nominal	$A \neq B$ represent different attribute manifestations
Ordinal	If $A \neq B$, then either $A < B$ or $A > B$
Interval	If $B - A = E - D$, then $B - A.$ and $E - D$ represent the same differences in the attribute
Ratio	If $\frac{A}{B} = 2$, then A represents twice the amount of the attribute as B (i.e., $A = 2B$)

Uniqueness

Scale Type	Allowable Transformations
Absolute	Identity
Nominal	Any that preserve equivalence classes (e.g., one-to-one transformations)
Ordinal	Monotonic increasing
Interval	Affine
Ratio	Rescaling

Uniqueness

Measurement Scales

Not up to instrument developers/users to determine the scale type

- Scale type determined by what is currently known about the relations that exist among the different attribute manifestations

Scaling

Definition (Scaling)

Scaling is the process of assigning numerals (or other symbols) to objects to reflect what is known about their manifestation of the attribute.

Scaling

Numerals should be selected in a way so that the values

- Are useful to those who have cause to measure the attribute
- Minimises any misinterpretations or unwarranted inferences about the attribute

Scaling

Definition (Measurement Unit)

A measurement unit is a particular quantity for the specified magnitude.

Scaling

- Counting is the basis of many scores from psychological instruments
- Counts $\neq$ units (Cooper & Humphry, 2012)

Scaling

Measurement Units

Additivity (i.e., interval scale) is required for measurement units to make sense.

Uncertainty

Definition (Measurement Uncertainty)

Measurement uncertainty is a value associated with the measurement result that characterises the range of reasonable values for the attribute manifestation (Joint Committee for Guides in Metrology, 2008).

Uncertainty

Measurement uncertainty components

- Imperfect objectivity
- Imperfect inter-subjectivity

Uncertainty

Definition (Measurement Objectivity)

If a measurement has perfect objectivity (object-dependence) then the measurement values we obtain are *solely* influenced by the attribute of interest in the measured objects (Ramsey et al., 2011).

Uncertainty

Detractors from objectivity

- Sampling (i.e., how well the sample selected to measure represents population)
- Object (i.e., issues surrounding the objects manifesting the attribute)
- Testing and measurement methods (i.e., issues surrounding the test conditions or instruments used)
- Measurement basis (i.e., inter-subjectivity issues).

Uncertainty

Measurement Objectivity

Impossible to remove *all* effects of external influences in measurement

Uncertainty

Need to estimate uncertainty when (Joint Committee for Guides in Metrology, 2008):

- Purpose of measurement: make "high-stakes" decisions
- Amount of uncertainty is not ignorable

Uncertainty

Definition (Measurement Inter-subjectivity)

If a measurement has perfect inter-subjectivity (subject-dependence) then the information acquired is not dependent on the individuals conducting the measurement or the particular instruments they use (Maul, Mari, & Wilson, 2019).

Uncertainty

Measurement Inter-subjectivity

Inter-subjectivity means the measurement results are unambiguously interpretable across difference circumstances.

Uncertainty

Standards (reference properties) for measuring many physical attributes

- Can trace back the measurement to some common reference

Psychological Measurement

Can We Measure Psychological Attributes?

Can psychological attributes be measured?

- Depends on who you ask

Can We Measure Psychological Attributes?

Received view

- Any attribute can be measured
- Just create a rule to assign numbers

Can We Measure Psychological Attributes?

Quantity view

- Only attributes that can be measured are quantitative (i.e., has both ordinality and additivity)
- Unless we can demonstrate an attribute in quantitative, it cannot be measured

Can We Measure Psychological Attributes?

Middle view

- Measurement: epistemic *process*, not a particular feature of the process' results or inputs (Mari, Maul, Irribarra, & Wilson, 2013)
- Not being quantitative does **not** disqualify us from potentially measuring it

Can We Measure Psychological Attributes?

Joint Committee for Guides in Metrology

- Only quantities can be measured with units
- **But**, allow for existence of *ordinal quantities*

Can We Measure Psychological Attributes?

Definition (Ordinal Quantity)

An ordinal quantity is the representation of an attribute whose manifestations can only be ordered (Joint Committee for Guides in Metrology, 2012).

Can We Measure Psychological Attributes?

Ordinal Quantity

- Cannot have measurement units
- Differences and ratios have no uniform meaning

Can We Measure Psychological Attributes?

What cannot be measured?

- Non-ordered categories (e.g., biological sex, country of origin, diagnosis)

Can We Measure Psychological Attributes?

What cannot be measured?

- Behaviour counts

Can We Measure Psychological Attributes?

Definition (Counts)

Counts are how many things are in a collection of similar things

Can We Measure Psychological Attributes?

Why aren't behaviour counts measurement?

- Each instance of the thing counted is treated as if it were interchangeable with every other instance
- Ignore whether
  - Manifestations are distinguishable (i.e., equivalent)
  - One manifestation is more of the attribute then another (i.e., ordinality).

Can We Measure Psychological Attributes?

Counts

- The more similar the counted things are (i.e,. homogeneous), the closer counting comes to measurement

Criterial Behaviours

Definition (Criterial Behaviours)

Criterial behaviours are behaviours that are constitutive of (i.e., follow the rules for) a psychological attribute's meaning (Baker & Hacker, 1982).

Criterial Behaviours

Criterial behaviours

- Establish attribute, **not** correlate with it (Witherspoon, 2011)
- Attribute-criteria relation is *grammatical* (i.e., philosophical), **not** empirical or even logical

Criterial Behaviours

Criterial behaviours

- Required when we want to communicate that another person has a particular psychological attribute

Criterial Behaviours

Measuring Psychological Attributes

To be able to measure a psychological attribute, there needs to be some core behaviours that are constitutive of the concept.

Some Classes of Psychological Attributes

Measuring psychological attributes requires mastering the concept

1. Invent concepts (neologisms)
2. Conceptual analysis of existing concepts

Some Classes of Psychological Attributes

Some classes of psychological attributes

1. *Ability*: general potentiality

Some Classes of Psychological Attributes

(Two-way) Ability

Having a (two-way) ability to do some thing is separate from doing the thing on a particular occasion

- Need:
  - Desire to exercise it
  - Opportunity to do so
  - Availability of necessary equipment

Some Classes of Psychological Attributes

Some classes of psychological attributes

2. *Beliefs*: what we believe to be so about something

Some Classes of Psychological Attributes

Some classes of psychological attributes

3. *Disposition*: more likely than not to do something, across multiple (but not all) circumstances

Some Classes of Psychological Attributes

Some classes of psychological attributes

4. *Passions*: (Hacker, 2018)
   - Affections (e.g., agitations, emotions, moods)
   - Appetites (e.g., hunger, thirst, sex)
   - Attitudes (e.g., sentiments, subjective value judgments)
   - Cogitative feelings (e.g., opinions, hunches)
   - Desires
   - etc.

Psychological Testing and Instrumentation

Variety of ways to collect information about behaviours

- "Best" method depends on nature of attribute

Psychological Testing and Instrumentation

Some behaviours only require observation

Psychological Testing and Instrumentation

Other behaviours are difficult to observe directly, but we can rely on avowals

- Interview
- Questionnaire

Psychological Testing and Instrumentation

Some behaviours do not lend themselves to observation or avowals

- To elicit necessary behaviours, have to use testing

Psychological Testing and Instrumentation

Definition (Testing)

Testing involves applying a set of procedures under particular environmental conditions in order to observe objects' attribute manifestations (Czichos, 2011).

Psychological Testing and Instrumentation

Definition (Psychological Testing)

Psychological testing is testing that designed to elicit a sample of behaviours

Psychological Testing and Instrumentation

Definition (Psychological Testing Items)

Psychological testing items are stimuli (e.g., statements, questions, tasks) designed to elicit certain behaviours.

Measurement Properties

Representation

- What do we know about the equivalence classes of psychological attributes?
  - Different classes represent distinction aspects of phenomenon?
  - Different classes represent more or less of the phenomenon? (ordinality)
  - Different classes can be combined to represent a distinct class? (additivity)

Measurement Properties

Representation

Without knowing the empirical structure of a psychological attribute, difficult to know how to represent it

Measurement Properties

Uniqueness

- Psychologists often work backwards by empirically studying what transformations are admissible and then determining the scale type
- Becomes problematic when interpreting score values

Measurement Properties

Scaling

- Three common scaling approaches for psychological attributes (Torgerson, 1958)
  - Object-centred
  - Stimulus-centred
  - Response-centred

Measurement Properties

Scaling Approaches in Psychological Measurement

Approach	Response Variation Attribution	Scaling Focus	Example
Stimulus	Stimuli's relation to attribute	Stimuli	Psychophysics
Object	Individual differences on attribute	Respondents	Classical test theory
Response	Individual differences on attribute & stimuli's relation to attribute	Respondents & stimuli	Rasch

Measurement Properties

Measurement Units

For psychological attributes, none of the scaling approaches necessarily provide a measurement unit.

- Possible exception: response time

Measurement Properties

Consequence of not having measurement unit

- Measurement values have to be "enhanced" to have meaning (Petersen, Kolen, & Hoover, 1989)

Uncertainties

Uncertainties in psychological measurement

- Testing procedures
- Measurement

Uncertainties

Definition (Testing Uncertainties)

Testing uncertainties are uncertainties in measurement results that arise due to adherence to the procedures for eliciting behaviour.

Uncertainties

Definition (Measurement Error)

Measurement error is the disagreement between the measurement result and the attribute manifestation.

Uncertainties

Inter-subjectivity

No standards (reference properties) for measuring most/all psychological attributes

- Cannot trace back the measurement to some common reference if there are concerns about measurement error

Uncertainties

Substitution

We have substituted

- *Measurement agreement*: agreement between the measurement result and the attribute manifestation
  - *intra*-individual variability

for

- *Reliability*: measurement consistency across multiple individuals
  - *inter*-individual variability

Test Scores

Scaling

Petersen, Kolen, & Hoover, 1989, p. 222

the main purpose of scaling is to aid users in interpreting test results. In this vein, we stress the importance of incorporating useful meaning into score scales as a primary means of enhancing score interpretability.

Raw Scores

Simplest scaling: raw scores

Raw Scores

Definition (Raw Score)

A raw score is the expression of some performance in terms of a particular scale's unit (Freeman, 1926).

Raw Scores

Psychological Attributes

Reality: most psychological attributes are either

- Qualitative
- Ordinal quantities

so cannot have measurement units

Raw Scores

Upshot

Current state of knowledge about most psychological attributes is such that they cannot have measurement units

Raw Scores

Raw Scores from Psychological Instruments

Most raw scores from psychological instruments are **behaviour counts**

- No meaning outside of a particular instrument

Raw Scores

What use are raw scores?

- Rank ordering performance

Raw Scores

Upshot

Counts can be clinically useful, but values are not measurements

- Counts: not directly comparable across instruments (or behaviour codes).

Raw Scores

Incorporating Meaning Into Psychological Instruments' Scores

Scores from psychological instruments need to be transformed to incorporate meaning.

Incorporating Normative Meaning

Definition (Norm-referenced Score)

A norm-referenced score is a scale score that incorporates normative information.

Incorporating Normative Meaning

Definition (Norm Group)

A norm group (standardization sample) is a sample of individuals used to establish normative behaviours for the population it represents.

Incorporating Normative Meaning

Definition (Scaled Scores)

Scaled scores are the resulting values after (re)scaling raw scores.

Incorporating Normative Meaning

Common ways to transform raw scores (R) into scale scores (S)

1. Linear (2 points of equivalence)
2. Non-linear (> 2 points of equivalence)

Incorporating Normative Meaning

Linear Transformation

Linear transformations take the form of

$$S = a + bR, \tag{1}$$

where

- a: intercept (location) of the line
- b: is the slope (spread).

Incorporating Normative Meaning

Nonlinear transformations can take variety of forms

- Percentiles
- Normalising scores (i.e., make scale scores follow normal distribution)

Incorporating Normative Meaning

Early approach to incorporating normative meaning: percentiles (Galton, 1885)

Incorporating Normative Meaning

Definition (Percentile)

A percentile (percentile point or centile) is the is value of a measurement/variable below which a specified percentage of a particular group's scores fall (Kirk, 2008)

Incorporating Normative Meaning

Robert S. Woodworth

- Standard score (or Z score)

Incorporating Normative Meaning

(Simple) Process of creating standard scores

1. Administer instrument to a norm group
2. Produce raw scores for everyone in norm group
3. Calculate raw score average
4. Calculate raw score dispersion
5. Calculate how many dispersion "units" each raw score is from average

Incorporating Normative Meaning

Standard Score

$$\text{Standard Score or } Z \text{ score} = \frac{\text{Raw} - \overline{\text{Raw}}}{D_{\text{Raw}}}, \tag{5}$$

where

- Raw: measurement value in its original *raw* metric for a specific person
- $\overline{\text{Raw}}$: average (e.g., mean, median) raw score value from norm group
- D_{Raw}: dispersion (e.g., standard deviation) of raw score value in norm group

Incorporating Normative Meaning

Raw-Score to Standard Score [Scaling Perspective]

$$S = \frac{R - \mu_{R(NG)}}{\sigma_{R(NG)}}$$

where

- $\mu_{R(NG)}$: mean of raw scores in norm group
- $\sigma_{R(NG)}$: SD of raw scores in norm group

Incorporating Normative Meaning

Z scores

Problem with Z scores

- Can have negative values
- Have to deal with non-integer values

Incorporating Normative Meaning

Linear Transformation of Z scores

$$S = \underbrace{\mu_S}_{a} + \underbrace{\sigma_S}_{b} Z_R, \tag{6}$$

where

- Z_R: Z-score transformation of the raw score value
- μ_N: desired (new) mean
- σ_N: desired (new) SD

Incorporating Normative Meaning

Common Means and SDs for Standard Score Transformations.

Scale	μ_N	μ_N
IQ	100	15
"Scaled Score" (Wechsler)	10	3
T (Thorndike)	50	10
Normal Curve Equivalent	50	21.06
Stanine	5	1.96
SAT	500	100
ACT	18	6
GRE (Verbal & Quantitative)	150	8.75

Incorporating Normative Meaning

Norm-referenced scores

- Useful to compare individual to population typicality
  - e.g., Diagnosis of osteoporosis (Miller, 2006)

Incorporating Normative Meaning

Information Loss

Creating norm-referenced scores looses information: meaningful relation between score and character of attribute score represents

Incorporating Normative Meaning

Upshot

Singular reliance on norm-referenced scores in interpreting psychological instruments is problematic.

Incorporating Content Meaning

Definition (Incorporating Content Information)

Incorporating content into scores involves placing information into scores related to functional skills or test content

Incorporating Content Meaning

Common way to incorporate contact information: "add ons" to describe norm-referenced scores

- Provide (arbitrary) qualitative descriptors of (arbitrary) numerals

Incorporating Content Meaning

- Is this useful (Woods et al., 2018)

Example (WISC-V Qualitative Descriptors)

IQ score	Qualitative Description
< 69	Extremely Low
$70 - 79$	Very Low
$80 - 89$	Low Average
$90 - 109$	Average
$110 - 119$	High Average
$120 - 29$	Very High
> 130	Extremely High

Incorporating Content Meaning

Better: Score values represent functional or meaningful aspects of attribute (Kolen & Brennan, 2014).

- Item mapping
- Scale anchoring
- Standard setting

Incorporating Content Meaning

Definition (Item Mapping)

Item mapping requires associating items with particular scale scores.

Incorporating Content Meaning

Definition (Scale Anchoring)

Scale anchoring provides general statements about what individuals with certain scores know or can do.

Incorporating Content Meaning

Scale anchoring: Elaboration of item mapping

- Find items that map around a particular score
- Review item content and develop general statements that represent skills at particular level

Incorporating Content Meaning

Definition (Standard Setting)

Standard setting involves finding a scale score that differentiates those who have and do not have some qualitative attribute

Incorporating Precision into Scores

Definition (Incorporating Precision)

Incorporating precision is the process of determining the number of values for an instrument's scores that will be available for use.

Incorporating Precision into Scores

Psychological instrument scores: no units

- Number of possible score values selected by instrument developer or publisher

Incorporating Precision into Scores

Variety of methods

- Most based on psychometric rules of thumb $\rightarrow$ miss the broader point

Incorporating Precision into Scores

Broader Point of Score Precision

Number of values from an instrument should be determined by the number of attribute equivalent classes known to exist.

Incorporating Precision into Scores

Ordinal and nominal attributes

- Number of values **should** be based on distinct levels of attribute known to exist

Assessing Growth

Measuring Attribute Change

Measuring attribute change can be problematic with norm-referenced and content-referenced scores.

Norm-referenced scores

- Do **not** assess change well
- Scores are ranks/relative positions: stable over long time periods for many psychological attributes

Change in Standard Scores

For an individual's standard score to increase, the individual must change at a **faster** rate than norm group.

Content-referenced scores

- Values: qualitative descriptors
- Scores often have to reach a certain threshold before change in category

Grade-Equivalent Scores

Definition (Grade Equivalent)

The grade equivalent of a particular test score, X, is the grade level for which X is the median value (Flanagan, 1951)

Grade-Equivalent Scores

General process (Petersen et al., 1989)

1. Administer instrument to students in desired grades
   - Instrument must cover content/skills for students in all desired grades
2. Calculate scores for each student on single scale (*interim-score scale*)
   - eg, score from fourth-grade student directly comparable to that from fifth-grade student
3. Rank order students on interim-score scale

Grade-Equivalent Scores

General process (Petersen et al., 1989)

5. Calculate interim-score values for intermediate grade levels (interpolating)
6. Calculate grade equivalent values from scaled interim-scores

Grade-Equivalent Scores

General process (Petersen et al., 1989)

7. Calculate interim-score values for "extramediate" (outside) grade levels

Grade-Equivalent Scores

General process (Petersen et al., 1989)

8. Calculate grade equivalent values from scaled interim-scores

Grade-Equivalent Scores

Grade-Equivalent Score Problem

The problem with grade equivalents is not the scores themselves, but their strong propensity for misinterpretation.

Grade-Equivalent Scores

Grade Equivalent Interpretation

GE scores do **not** indicate:

- Where student should be placed in the graded organisation of school
- Whether a student has the minimum academic skill set for a specific grade

Questionable Practices

Assuming Score Exchangeability

Definition (Common-or-garden Concepts)

Common-or-garden concepts are those taught, learned and understood by the *person on the street*, and have meanings that are manifest in broad, normative linguistic practices.

Assuming Score Exchangeability

Definition (Technical Concepts)

Technical concepts are those defined by a specialised or expert community, and employed within a narrow, technical field of application. Usually, go beyond verbal definitions.

Assuming Score Exchangeability

Problem with Non-Technical Terms

A major problem with using non-technical terms to describe attributes is that they are apt cause confusion in scientific investigation and psychological measurement.

Assuming Score Exchangeability

Barrett, 2011, p. 29

The very real problem we face as psychological scientists is how to conceive of defining any psychological attribute in a clear, technical manner, such that we can propose experimental manipulations that might test our expectations about magnitude relations.

Score Conversions

Some score interpretation programs require standard scores on IQ scale (mean: 100, SD: 15)

- Purpose: directly comparable

Score Conversions

Score Transformation Formula

$$S_{IQ} = 100 + 15 \times \frac{S_O - M_O}{SD_O},$$

where

- O: original score scale
- S: student's score
- M: mean (norm group)
- SD: standard deviation (norm group)

Score Conversions

Linear transformations permissible for scores on interval & ordinal scales

Score Conversions

Interval scales: Quantitative

- No loss/gain of information by going from one scale to another via linear transformation
- Scale chosen: most convenient for given application (e.g., meter vs. inch)

Score Conversions

Linear transformation applied to ordinal scale values $\rightarrow$ only preserves rank orders

- **Only** interpretations related to scores being $\geq$ other scores are invariant

IQ Nonsense

g: Technical concept

- Spearman: Developed technical definition
- Anybody who reads his work knows exactly what he was referring when he used the term

IQ Nonsense

Spearman

- Abhorred the words “intelligence”
  - **Not** technical term

IQ Nonsense

Psychologists tried to map “ intelligence” onto g with less-than-useful results

IQ Nonsense

IQ

$$IQ = \sum_i a_i \text{subtest}_i$$

where

- Subtest: instrument purporting to assess some specific attribute within the intelligence sphere
- a: some weight of importance of a particular subtest (often $a = 1$)

IQ Nonsense

IQ value meaning differs from instrument to instrument

- Every IQ test author has own ideas about what specific attributes are important

IQ Nonsense

IQ taken as measurement of "intelligence"

- IQ has **no uniform** meaning—instrument specific
- "Intelligence" has no technical meaning

IQ Nonsense

Upshot

Unless the IQ subcomponents are the exact same across two instruments & and the instruments scores have been shown to be (or made to be) equivalent, why should be expect scores to be the same?

Going Forward

What to Do?

Psychological testing **can** be useful

- Need: realistic perspective

What to Do?

Score Realism

Most scores from psychological instruments represent some type of fuzzy order

What to Do?

Assess orders fairly well, especially in cognitive performance and knowledge domains

What to Do?

Classify fairly well

What to Do?

Score Realism, Redux

We **should not** pretend that we have better measurement than we actually have.

What to Do?

Upshot

Pretending numbers from psychological instruments mean something more than they do produces a false sense of accuracy in our decision-making.

What to Do?

Example (False Sense of Accuracy)

- "Significant" score differences map onto differences in attribute levels
- Difference in, e.g., 5 or 10 points across instruments represents different levels of an attribute
- "Quantitive" profiles of scores can accurately classify individuals

Questions?

Questions?

References & Suggested Readings

Baker, G. P., & Hacker, P. M. S. (1982). The grammar of psychology: Wittgenstein's *Bemerkungen Über die Philosophie der Psychologie*. *Language & Communication*, *2*, 227-244.

Barrett, P. (2011). Invoking arbitrary units is not a solution to the problem of quantification in the social sciences. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 28-31.

Barrett, P. T. (2001). *The role of a concatenation unit .* Paper presented at the annual meeting of the British Psychological Society, Mathematics, Statistics, and Computing Section, London, England.

International Reading Association. (1982). Misuse of grade equivalents: Resolution passed by the Delegates Assembly of the IRA, April 1981. *Reading Teacher*, *35*, 464.

References & Suggested Readings (cont.)

Joint Committee for Guides in Metrology. (2008). *JCGM 100:2008. Evaluation of measurement data—Guide to the expression of uncertainty in measurement*. Author. Retrieved from https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

Joint Committee for Guides in Metrology. (2012). *JCGM 200:2012. International vocabulary of metrology—basic and general concepts and associated terms (VIM)* (3rd ed.). Sévres, France: Author. Retrieved from https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf.

Binet, A. (1910). Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école. *L'Année psychologique*, *17*, 145-201.

Capron, C., Vetta, A. R., Duyme, M., & Vetta, A. (1999). Misconceptions of biometrical IQists. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *18*, 115-160.

References & Suggested Readings (cont.)

Cepeda, J., Nicholas, Blackwell, A., Katharine, & Munakata, Y. (2013). Speed isn't everything: Complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science*, *16*, 269-286.

Cooper, G., & Humphry, S. M. (2012). The ontological distinction between units and entities. *Synthese*, *187*, 393-401.

Czichos, H. (2011). Introduction to metrology and testing. In H. Czichos, T. Saito, & L. Smith (Eds.), *Springer handbook of metrology and testing* (pp. 3–22). Berlin, Heidelberg: Springer. Retrieved from https://doi.org/10.1007/978-3-642-16641-9_1.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22*, 15-25.

Emerson, W. H. (2004). One as a 'unit' in expressing the magnitudes of quantities. *Metrologia*, *41*, L26-L28.

References & Suggested Readings (cont.)

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Freeman, F. N. (1926). *Mental tests: Their history, principles and applications*. Boston, MA: Houghton, Mifflin and Company.

Galton, F. (1885). Some results of the Anthropometric Laboratory. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *14*, 275-287.

Hacker, P. M. S. (2018). *The passions: A study of human nature*. New York, NY: Wiley.

Jaffe, L. E. (2009). *Development, interpretation, and application of the W score and the relative proficiency index* (Tech. Rep. No. Woodcock-Johnson III Assessment Service Bulletin 11). Rolling Meadows, IL: Riverside Publishing.

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Thomson Wadsworth.

References & Suggested Readings (cont.)

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., p. 155-186). Westport, CT: American Council on Education/Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology*, *8*, 435-461.

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, *31*, 32-42.

References & Suggested Readings (cont.)

Mari, L., Carbone, P., & Petri, D. (2015). Fundamentals of hard and soft measurement. In A. Ferrero, D. Petri, P. Carbone, & M. Catelani (Eds.), *Modern measurements: Fundamentals and applications* (pp. 203–262). Hoboken, NJ: Wiley-IEEE Press. Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7304075.

Mari, L., Maul, A., Irribarra, D. T., & Wilson, M. (2013). Quantification is neither necessary nor sufficient for measurement. *Journal of Physics: Conference Series*, *459*, 1-6.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764-770.

Miller, P. D. (2006). Guidelines for the diagnosis of osteoporosis: T-scores vs fractures. *Reviews in Endocrine and Metabolic Disorders*, *7*, 75-89.

References & Suggested Readings (cont.)

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903.

Ramsey, M. H., Ellison, S. L. R., Czichos, H., Hässelbarth, W., Ischi, H., Wegscheider, W., ... Steiger, T. (2011). Quality in measurement and testing. In H. Czichos, T. Saito, & L. Smith (Eds.), *Springer handbook of metrology and testing* (pp. 39–141). Berlin, Heidelberg: Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

References & Suggested Readings (cont.)

Texas State Board of Examiners of Psychologists. (2019). *Texas State Board of Examiners of Psychologists acts and rules*. Austin, TX: Author.

Thorndike, E. L. (1918a). Individual differences. *Psychological Bulletin*, *15*, 148-159.

Thorndike, E. L. (1918b). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The Seventeenth Yearbook of the National Society for the Study of Education* (Vol. 2, pp. 16–24). Bloomington, IL: Public School Publishing.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

Witherspoon, E. (2011). Wittgenstein on criteria and the problem of other minds. In O. Kuusela & M. McGinn (Eds.), *The Oxford handbook of Wittgenstein* (pp. 472–498). New York, NY: Oxford University Press.

References & Suggested Readings (cont.)

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 105–127). Mahwah, NJ: Erlbaum.

Woods, I. L., Floyd, R. G., Singh, L. J., Layton, H. K., Norfolk, P. A., & Farmer, R. L. (2018). What is in a name? A historical review of intelligence test score labels. *Journal of Psychoeducational Assessment*, 0734282918786651.

Woodworth, R. S. (1912). Combining the results of several tests: A study in statistical method. *Psychological Review*, *19*, 97-123.