A. Alexander Beaujean
Associate Professor of Psychology & Neuroscience

Unmuddled Measurement
Straight talk about psychological measurement and score interpretation
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”
A quantitative phenomenon is one where the relations among distinguishable elements have both order and additivity.
Before representing attributes we need empirical (experimental) investigations
Psychological attributes

- Any empirical reason to believe they are quantitative?
- Or are these just qualitative distinctions?
Can use numbers to represent things

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

Upshot

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

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

Goal of measurement: provide *information* about a specific *attribute* that describes the empirical world effectively.
Concepts are elements of language used to denote and organise phenomena (Maraun & Gabriel, 2013).
Concepts

Concepts are human creations/inventions.
Concept meaning

- Fixed by linguistic rules
  - Determine its range of correct employments
- Meaning manifested: range of correct employments
Measurement concepts of interest: object attributes
Clarify concepts and their constitutive referents before beginning measurement
Attributes: not directly measurable (they are concepts)

- Instead, measure attribute *manifestations*
Attribute manifestations: divide into classes that are **mutually exclusive** and **exhaustive**

- **Classes**
  - Do **not** overlap
  - Contain all possible manifestation of attribute
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 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)
Discover information about attribute manifestation relations ⇒ 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
Attribute manifestations need to meet certain conditions to be measurable

- Representation
- Uniqueness
- Meaningfulness
- Scaling
- (Uncertainty)
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)
A measurement **scale** is the particular way we assign numbers to measure the attribute.
Uniqueness

Typically $\geq 1$ set of symbols we can use

- More empirically-discovered relations to preserved $\Rightarrow$ fewer unique symbol systems available
Uniqueness: closely related to value transformation

- Any symbol transformations we make must keep the representation intact (i.e., admissible transformation)
<table>
<thead>
<tr>
<th>Scale Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
<td>Counts</td>
</tr>
<tr>
<td>Nominal</td>
<td>$A \neq B$ represent different attribute manifestations</td>
</tr>
<tr>
<td>Ordinal</td>
<td>If $A \neq B$, then either $A &lt; B$ or $A &gt; B$</td>
</tr>
<tr>
<td>Interval</td>
<td>If $B - A = E - D$, then $B - A$ and $E - D$ represent the same differences in the attribute</td>
</tr>
<tr>
<td>Ratio</td>
<td>If $\frac{A}{B} = 2$, then $A$ represents twice the amount of the attribute as $B$ (i.e., $A = 2B$)</td>
</tr>
<tr>
<td>Scale Type</td>
<td>Allowable Transformations</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Absolute</td>
<td>Identity</td>
</tr>
<tr>
<td>Nominal</td>
<td>Any that preserve equivalence classes (e.g., one-to-one transformations)</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Monotonic increasing</td>
</tr>
<tr>
<td>Interval</td>
<td>Affine</td>
</tr>
<tr>
<td>Ratio</td>
<td>Rescaling</td>
</tr>
</tbody>
</table>
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 is the process of assigning numerals (or other symbols) to objects to reflect what is known about their manifestation of the attribute.
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
A measurement unit is a particular quantity for the specified magnitude.
Counting is the basis of many scores from psychological instruments.

Counts $\neq$ units (Cooper & Humphry, 2012)
Additivity (i.e., interval scale) is required for measurement units to make sense.
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).
Measurement uncertainty components

- Imperfect objectivity
- Imperfect inter-subjectivity
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).
Measurement Objectivity

Impossible to remove *all* effects of external influences in measurement
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
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).
Measurement Inter-subjectivity

*Inter-subjectivity* means the measurement results are unambiguously interpretable across different circumstances.
Standards (reference properties) for measuring many physical attributes

- Can trace back the measurement to some common reference
Psychological Measurement
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*
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 than 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 are behaviours that are constitutive of (i.e., follow the rules for) a psychological attribute’s meaning (Baker & Hacker, 1982).
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

- 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.
Measuring psychological attributes requires mastering the concept

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

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

   - 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.
Variety of ways to collect information about behaviours

▶ “Best” method depends on nature of attribute
Some behaviours only require observation
Other behaviours are difficult to observe directly, but we can rely on avowals

- Interview
- Questionnaire
Some behaviours do not lend themselves to observation or avowals

- To elicit necessary behaviours, have to use testing
Definition (Testing)

Testing involves applying a set of procedures under particular environmental conditions in order to observe objects’ attribute manifestations (Czichos, 2011).
Psychological testing is testing that is designed to elicit a sample of behaviours.
Definition (Psychological Testing Items)

Psychological testing items are stimuli (e.g., statements, questions, tasks) designed to elicit certain behaviours.
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)
Without knowing the empirical structure of a psychological attribute, difficult to know how to represent it
Uniqueness

- Psychologists often work backwards by empirically studying what transformations are admissible and then determining the scale type
- Becomes problematic when interpreting score values
Three common scaling approaches for psychological attributes (Torgerson, 1958)

- Object-centred
- Stimulus-centred
- Response-centred
## Measurement Properties

### Scaling Approaches in Psychological Measurement

<table>
<thead>
<tr>
<th>Approach</th>
<th>Response Variation Attribution</th>
<th>Scaling Focus</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td>Stimuli’s relation to attribute</td>
<td>Stimuli</td>
<td>Psychophysics</td>
</tr>
<tr>
<td>Object</td>
<td>Individual differences on attribute</td>
<td>Respondents</td>
<td>Classical test theory</td>
</tr>
<tr>
<td>Response</td>
<td>Individual differences on attribute &amp; stimuli’s relation to attribute</td>
<td>Respondents &amp; stimuli</td>
<td>Rasch</td>
</tr>
</tbody>
</table>
Measurement Properties

Measurement Units

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

- Possible exception: response time
Consequence of not having measurement unit

- Measurement values have to be “enhanced” to have meaning (Petersen, Kolen, & Hoover, 1989)
Uncertainties in psychological measurement

- Testing procedures
- Measurement
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.
### 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
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
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.
Simplest scaling: raw scores
Definition (Raw Score)

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

Reality: most psychological attributes are either

- Qualitative
- Ordinal quantities

so cannot have measurement units
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
What use are raw scores?

- Rank ordering performance
Upshot

Counts can be clinically useful, but values are not measurements

- Counts: not directly comparable across instruments (or behaviour codes).
Scores from psychological instruments need to be transformed to incorporate meaning.
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.
Definition (Scaled Scores)

*Scaled scores* are the resulting values after (re)scaleding raw scores.
Common ways to transform raw scores ($R$) into scale scores ($S$)

1. Linear (2 points of equivalence)
2. Non-linear ($> 2$ points of equivalence)
Linear transformations take the form of

\[ S = a + bR, \]  \hspace{1cm} (1)

where

- \( a \): intercept (location) of the line
- \( b \): is the slope (spread).
Nonlinear transformations can take variety of forms

- Percentiles
- Normalising scores (i.e., make scale scores follow normal distribution)
Early approach to incorporating normative meaning: percentiles (Galton, 1885)
Incorporating Normative Meaning

Definition (Percentile)

A percentile (percentile point or centile) is the 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)
(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

Standard Score or $Z$ score = \( \frac{\text{Raw} - \overline{\text{Raw}}}{D_{\text{Raw}}} \), \hspace{1cm} (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
- **\( D_{\text{Raw}} \)**: 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
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

\[ S = \mu_S \left[ \frac{a}{Z_R} \right] + \sigma_S \left[ \frac{b}{Z_R} \right] \]  

(6)

where

- \( Z_R \): Z-score transformation of the raw score value
- \( \mu_N \): desired (new) mean
- \( \sigma_N \): desired (new) SD
Common Means and SDs for Standard Score Transformations.

<table>
<thead>
<tr>
<th>Scale</th>
<th>$\mu_N$</th>
<th>$\mu_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>“Scaled Score” (Wechsler)</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>$T$ (Thorndike)</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Normal Curve Equivalent</td>
<td>50</td>
<td>21.06</td>
</tr>
<tr>
<td>Stanine</td>
<td>5</td>
<td>1.96</td>
</tr>
<tr>
<td>SAT</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>ACT</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>GRE (Verbal &amp; Quantitative)</td>
<td>150</td>
<td>8.75</td>
</tr>
</tbody>
</table>
Norm-referenced scores

- Useful to compare individual to population typicality
  - e.g., Diagnosis of osteoporosis (Miller, 2006)
Creating norm-referenced scores loses 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 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
<table>
<thead>
<tr>
<th>IQ score</th>
<th>Qualitative Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 69</td>
<td>Extremely Low</td>
</tr>
<tr>
<td>70 – 79</td>
<td>Very Low</td>
</tr>
<tr>
<td>80 – 89</td>
<td>Low Average</td>
</tr>
<tr>
<td>90 – 109</td>
<td>Average</td>
</tr>
<tr>
<td>110 – 119</td>
<td>High Average</td>
</tr>
<tr>
<td>120 – 29</td>
<td>Very High</td>
</tr>
<tr>
<td>&gt; 130</td>
<td>Extremely High</td>
</tr>
</tbody>
</table>
Better: Score values represent functional or meaningful aspects of attribute (Kolen & Brennan, 2014).

- Item mapping
- Scale anchoring
- Standard setting
Definition (Item Mapping)

Item mapping requires associating items with particular scale scores.
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
Definition (Standard Setting)

**Standard setting** involves finding a scale score that differentiates those who have and do not have some qualitative attribute.
Incorporating precision is the process of determining the number of values for an instrument’s scores that will be available for use.
Psychological instrument scores: no units

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

- Most based on psychometric rules of thumb → 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.
Ordinal and nominal attributes

- Number of values **should** be based on distinct levels of attribute known to exist
Assessing Growth
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
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
The grade equivalent of a particular test score, $X$, is the grade level for which $X$ is the median value (Flanagan, 1951).
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
General process (Petersen et al., 1989)

7. Calculate interim-score values for “extramediante” (outside) grade levels
General process (Petersen et al., 1989)

8. Calculate grade equivalent values from scaled interim-scores
The problem with grade equivalents is not the scores themselves, but their strong propensity for misinterpretation.
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.
**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.
A major problem with using non-technical terms to describe attributes is that they are apt cause confusion in scientific investigation and psychological measurement.
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.
Some score interpretation programs require standard scores on IQ scale (mean: 100, SD: 15)

- Purpose: directly comparable
Score Transformations

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
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)
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
Spearman

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

Psychologists tried to map "intelligence" onto $g$ with less-than-useful results
IQ Nonsense

\[ IQ = \sum_i a_{\text{subtest}_i} \]

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 value meaning differs from instrument to instrument

- Every IQ test author has own ideas about what specific attributes are important
IQ taken as measurement of “intelligence”

- IQ has **no uniform** meaning—instrument specific
- “Intelligence” has no technical meaning
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
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
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?


