Introduction to psychometrics

Christophe Lalanne

Fall 2011
## Outline

<table>
<thead>
<tr>
<th>What is psychometrics?</th>
<th>Items and scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key concepts</td>
<td>Measurement models</td>
</tr>
<tr>
<td>Measurement and instrument</td>
<td>Reliability and validity</td>
</tr>
<tr>
<td>Latent variable models</td>
<td></td>
</tr>
</tbody>
</table>

“It is rather surprising that systematic studies of human abilities were not undertaken until the second half of the last century. . . An accurate method was available for measuring the circumference of the earth 2,000 years before the first systematic measures of human ability were developed.” —Nunnally and Bernstein, 1994
What is psychometrics?

Psychometrics has evolved as a subfield of psychology to become the science of measurement of unobservable individual characteristics. It encompasses

• the construction and analysis of measurement instruments;

• the development of theoretical approaches to measurement.

It is sometimes opposed to clinimetrics which encourages the use of clinical expertise instead of formal statistical models, or even biometrics (de Vet et al., 2003).
“If Foo is a science then Foo often has both an area Foometrics and an area Mathematical Foo. Mathematical Foo applies mathematical modeling to the Foo subject area, while Foometrics develops and studies data analysis techniques for empirical data collected in Foo. Each of the social and behavioural sciences has a form of Foometrics, although they may not all use a name in this family.” (de Leeuw, 2006)
Some landmarks

G. T. Fechner
(1801–1887)
W. Wundt
(1832–1920)
G. von Békésy
(1899–1972)
R. Likert
(1903–1981)
E. H. Weber
(1795–1878)
S. S. Stevens
(1906–1973)
L. L. Thurstone
(1887–1955)
F. Galton
(1822–1911)
K. Pearson
(1857–1936)
L. L. Thurstone
(1887–1955)
L. J. Cronbach
(1916–2001)
A. Jensen
(1923–)
J. McKeen Cattell
(1860–1944)
C. Spearman
(1863–1945)
G. Rasch
(1901–1980)
F. M. Lord
(1912–2000)
Key concepts

We, human beings, like to make either relative or absolute judgments. Although it is often easy to tell whether a given object is taller than another object, it is harder to tell whether someone is more proficient in a given discipline than another person. Likewise, it is difficult to assess groupwise opinions by asking a random set of questions. Assessing one’s preference for a set of objects is also a difficult task. Why?

Mostly because the preceding points call for a process that allows to rate and scale the above object of measurement. Moreover, judgments elicited this way will rather lead to subjective measures, unlike most physical measurements.
never — almost never ≡ always — almost always

Ultimately, we may want to build a measurement tool that shares the same characteristics as the ones shown below, or at least one that allows us to locate and compare individuals’ responses or objects’ properties using some metrical system.
Measurement and instrument

An instrument is used to relate, or ‘map’, something observed in the real world to some other thing measured as part of a theory. The former is called a manifest variable, while the latter is usually subsumed under the term latent variable or factor.

Measurement is best understood as the task of assigning number to categories (Stevens, 1946), in order for the researcher to “provide a reasonable and consistent way to summarize the responses that people make to express their achievements, attitudes, or personal points of view through instruments such as attitude scales, achievement tests, questionnaires, surveys, and psychological scales.” (Wilson, 2005)
Latent variable models

A good overview of the use of LVMs in biomedical research is provided by Rabe-Hesketh and Skrondal, 2008.

For a more general framework, we can consider how latent and manifest variables relate each other: (Bartholomew and Knott, 2011)

<table>
<thead>
<tr>
<th>Latent variables</th>
<th>Manifest variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrical</td>
<td>Metrical: Factor analysis</td>
</tr>
<tr>
<td>Categorical</td>
<td>Latent profile analysis</td>
</tr>
</tbody>
</table>
The four building blocks (Wilson, 2005)

- A **Construct map** features a coherent and substantive definition for the content of the construct which is composed of an underlying continuum (for ordering respondents and/or items responses).

- **Items design** deals with the standardized construction of items that are supposed to stimulate responses, assimilable to observations about the construct.

- The **Outcome space** is the set of well-defined categories, finite and exhaustive, ordered, context-specific, and research-based.

- A **Measurement model** is needed in order to relate the scored outcomes from the items design and the outcome space back to the construct that was the original inspiration of the items.
Self-reported measures
What is an item?

An item constitutes the fundamental unit of measurement, and it can be considered as a criteried way to measure a psychological construct otherwise unobservable (e.g. reading fluency, specific cognitive ability, quality of life, personality traits).

Example: Below are nine items from the MOS 36-item short form, scored 1 (‘All of the Time’) to 6 (‘None of the Time’) points.

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>23. Did you feel full of pep?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>24. Have you been a very nervous person?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>25. Have you felt so down in the dumps that nothing could cheer you up?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>26. Have you felt calm and peaceful?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>27. Did you have a lot of energy?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>28. Have you felt downhearted and blue?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>29. Did you feel worn out?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>30. Have you been a happy person?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>31. Did you feel tired?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
From items to scale

Aggregating items together leads to a scale, which is often confounded with the dimension it is assumed to reflect. Provided it only measures a single construct (unidimensionality), we can assign scores to patterns of observed responses (e.g., summated-scale score) to infer one’s location on the underlying latent trait, but more generally rank items or individuals on this scale.

Two desired properties of a scale are unidimensionality and local independence. That is,

- The instrument should measure a single construct which accounts for one’s response to a specific item given his location on the latent trait (and not other factors);
If the latent variable is maintained at a given level, all the item responses should be independent.

Quoting Falissard, 2006, “A set of items is unidimensional if there exists a variable (often called a latent variable, as this variable may not be observed) which ‘explains’ all the correlations observed between the items”, but conceiving a single construct as underlying a broad concept (e.g., depression) might be misleading, see e.g. (Borsboom, 2006). Factor analytic methods can be used to study the dimensionality of an instrument, and provide much better information than single-number summary which are often inappropriately used (e.g., Cronbach’s alpha). Local independence is required in many models, including IRT or LCA models.
Examples of multidimensional instruments

- The **HADS** includes two scales (\(2 \times 7\) items, 21 points max.) for exploring anxiety and depression-related symptoms (Zigmond and Snaith, 1983 and Herrmann, 1997).
- The **NEOPI-R** is a personality inventory (240 or 60 items) relying on the five-factor model (extraversion, agreeableness, conscientiousness, neuroticism, and openness; 6 facets) (McCrae and John, 1992 and McCrae and Costa, 2004).
- The **MOS-HIV** for HIV/AIDS includes 35 items assessing: general health perceptions, pain, physical functioning, role functioning, social functioning, energy/fatigue, mental health, health distress, cognitive function, and quality of life (Wu et al., 1997a, 1997b).

A list of existing HRQL instruments is available on [http://www.proqolid.org/](http://www.proqolid.org/).
Two common measurement models

The classical test theory assumes that there exists a true score, defined as the expected response of an individual following repeated independent administration of the same test (error-free measure). But we only have access to

\[
\text{observed score} = \text{true score} + \text{some error}
\]

Although it has long been considered as a principled method in measurement, it suffers from different flaws, well reviewed in (Borsboom, 2006). Connected conceptual issues are reliability of test scores and item analysis. Here, a respondent having a ‘higher amount of the latent trait’ will have a higher score, usually computed as the sum of his responses.
The CTT approach includes few but simple mathematical formulation, with straightforward estimation of model parameters; it also yields scoring rules of practical interest.

**IRT models** on the contrary rely on a mathematical model whereby the probability of endorsing an item depends on both item and person location (parameters) on the latent trait. For every item in a scale, a set of properties (item parameters) are estimated. The item slope or discrimination parameter describes how well the item performs in the scale in terms of the strength of the relationship between the item and the scale. The item difficulty or threshold parameter(s) identifies the location along the construct’s latent continuum where the item best discriminates among individuals.
A well-known example is the Rasch model which deals with binary responses and assumes that all items are equally discriminative. IRT models are more flexible and do not assume an fixed axiomatically fixed relationship between the observed and true score. They allow for the joint scaling of person and item parameters when mapping the relationship between latent traits and responses to test items.

The use of IRT model is particularly interesting when the objective is to develop bank of calibrated items (Reeve et al., 2007), to deliver targeted and adaptive testing measures (Rebollo et al., 2010), or to study measurement invariance across groups of patients (Teresi, 2006).
Another illustration

The following example is taken from Partchev, 2004. Let us consider an item like

What is the area of a circle having a radius of 3 cm?

1. 9.00 cm$^2$
2. 18.85 cm$^2$
3. 28.27 cm$^2$

The first of the three options is possibly the most naive one, the second is wrong but implies more advanced knowledge (the area is confused with the circumference), and the third is the correct one.
A probabilistic account of item responses
COSMIN Taxonomy (Mokkink et al., 2010)

Reliability
- Internal consistency
- Reliability^a
- Measurement error^a

Validity
- Content validity
- Face validity
- Criterion validity^b
- Structural validity
- Hypotheses testing
- Cross-cultural validity

Responsiveness
- Responsiveness

Interpretability

^a (test-retest, inter-rater, intra-rater)
^b (concurrent validity, predictive validity)
The many facets of validity

“Validation of instruments is the process of determining whether there are grounds for believing that the instrument measures what it is intended to measure, and that it is useful for its intended purpose. (Fayers and Machin, 2000)”

Let’s consider the following definitions: (Falissard, 2008)

- **Content validity** reflects the adequacy of the domains or dimensions spanned by the items;
- **Criterion validity** demonstrates that scales have empirical association with external criteria, such as gold standards or other instruments purported to measure equivalent concepts;
- **Construct validity** relates each inter-items and item-scale relationships from a theoretical point of view.
Convergent and discriminant validities are both subsumed in the general and theoretical concept of construct validity. A critical appraisal of the concept of validity can be found in (Zumbo, 2007 and Borsboom et al., 2004, 2009).

A questionnaire is valid to the extent test scores are interpreted in relation to the construct that the questionnaire purports to assess. An alternative (and more elegant) definition of construct validity is that it “is a property of measurement instruments that codes whether these instruments are sensitive to variation in a targeted attribute” (Borsboom et al., 2009).
Some caveats about clinical validity

In clinical setting, the validity of a medical diagnosis requires a clear aetiology. However, most of functional diagnosis of mental health-related disorders (e.g. personality disorder, schizophrenia) are defined in a circular fashion: The diagnosis is made on the basis of symptoms and the symptoms are accounted for by the diagnosis.

Likewise, although most psychiatrists will agree on what depression is, demonstrating categorically that clinical depression differs from dysphoria or everyday unhappiness is nearly impossible (Pilgrim, 2009).

Finally, the problems of valid and reliable case identification in psychiatric epidemiology remain unsolved because no physical cause are generally associated to the disease under consideration.
Different types of reliability

The sources of scores variability may be very different, e.g. (Dunn, 2000)

- repeated assessment of several subjects by the same rater or clinician;
- alternative assessment of a given subject by different raters;
- alternative administration of the same questionnaire or a parallel form;
- use of different subscales of a single questionnaire to infer one’s performance;

Reliability analysis aims at quantifying these variations, in order to provide an idea of the error of measurement.
Reliability vs. significance

A note of caution: One must distinguish between reliability and significance: (Thompson, 2003)

- ‘statistical’ significance evaluates the probability or likelihood of the sample results, with reference to a reference population where the null is exactly true; ‘practical’ and ‘clinical’ significance are closely related concepts underlying the extent to which sample results diverge from the null (as measured by effect size statistics) or the way treated patients may not be distinguished from control or normal ones.

- However, ‘statistical’, ‘practical’, and ‘clinical’ significance all stand on the assumption that scores are meaningful and reliable indicators of individuals’ performance...
Construction of a questionnaire

correspondence with RCT

phase II phase III phase IV

validation

content validity

construct validity concourant validity scores reliability

responsiveness face validity

test construction test administration follow-up

field testing large-scale testing

phase construction test administration follow-up


