

# Measurement, validity and reliability

*Chris Wild, University of Auckland*

Good measurement and classification is absolutely critical to good data, but it is not something statisticians tend to say much about.

Why? Because statistics tends to concentrate on places where statisticians can devise methods that can be very widely applied. It tends to concentrate on things which are generic. Measurement is inherently specific.

The people who know how to measure something well are the people who know it intimately, be it aspects of the way your body functions, the strengths of materials, aspects of the performance of companies, aspects of peoples' psychology, or the ways societies function. The people who think deeply about, and then learn how to measure these things, are the researchers in biology and medicine, engineering, accounting and economics, sociology, and so on.

You will often hear the names of two generic big ideas in measurement, however, "validity" and "reliability". A measure is *valid* if it is "measuring the right thing". It is *reliable* if, when you measure the same thing over and over again, you get pretty much the same answer. Clearly, however, there is not much point in your measure reliably giving the same answers if it is not measuring the right thing (it is not "valid"). But reliability is much easier to demonstrate than validity.

Here is a [series of questions](#) targeting some of the things you should think about if you are needing to devise ways to measure or classify something (from my [Statistical Thinking](#) images page, and [here](#)).

*Our story continues ...*

## Measurement, Reliability and Validity

By *Professor Jim Ridgway, The Smart Centre, Durham University*

The invention of useful measures has had a profound effect on human societies. Think about the introduction of standard measures (say of weight) on early trade, and the invention of a reliable measure of longitude on navigation (and all that followed). It is easy to think that the measures we use are 'obvious' - temperature, mass, time – but

each one has a history of conceptual and technical innovations. So what makes for a good measure?

Try your hand at inventing a measure of 'squareness'. You already know what the 'squareness' of a rectangle is – skinny ones are less 'square' than tubby ones. Here is a [link to a problem statement, and a solution](#).

The activity shows that the creation of a measure raises issues such as: validity – e.g. given a collection of rectangles, does the measure actually rank them on 'squareness' in the same way that human judgement does?; reliability – e.g. how stable is the measure if it used by different people?; practicability – e.g. how easy is it to use?

Both reliability and validity have to be judged in the context of the decisions that are going to be made. Reliability and validity are actually very tricky concepts, especially in the social and human sciences. Reliability is the easier of the two – if you repeat a measure in circumstances where you expect to get the same result, are the scores close enough for your purposes? Does my sofa fit in a furniture van? Does my sofa fit in my estate car? Arm span is a perfectly reliable measure of the sofa for the first question, but probably not for the second one.

Evaluating the validity of a measure requires an overall judgement of the extent to which the story about what is being measured (and supporting evidence) justifies the actions and interpretations that will be based on the measure. This should include a consideration of the social consequences of adopting a particular measure (e.g. using 'scholastic aptitude' for university admission, rather than subject specific knowledge). However, making an overall judgement of 'validity' does not mean that components (such as construct validity, concurrent validity, predictive validity...) can be ignored. Some authors see the validation of a measure as a 'one-off'; others argue that validation should be an on-going process. It is a brave (or foolhardy) person who believes that measures of any social phenomenon are stable over time.

For a comprehensive account of issues around measurement, with sensible advice about assessment in education and psychology, see Newton and Shaw (2014).

Newton, P. ,and Shaw, S. (2014) *Validity in Educational and Psychological Assessment*. Sage: London. ebook ISBN 9781473904064; paperback ISBN 9781446253236

© 2014 Chris Wild, The University of Auckland and Jim Ridgway, Durham University