Statistical methods used in the development of a health measurement scale

Pamela Warner

Reader in Medical Statistics, Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Correspondence to
Dr Pamela Warner, Centre for Population Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK;
p.warner@ed.ac.uk

Received 15 May 2012
Accepted 17 May 2012

Background

Various psychometric statistical methods have been used in a paper by Simon *et al*[1] in this issue of the Journal. These notes are intended to provide some additional explanation of the methods employed in developing a health measurement scale. [See Box 1 for a glossary of terms used in this article.]

What statistical methods are used in scale development?

A health measurement scale is a tool designed for a particular purpose: to quantify some attitude (say 'acceptability' of sterilisation), or to screen for those with a high-risk status (say, practising unsafe sex) in order to offer them additional counselling, or to assess cancer knowledge (as in the Simon *et al.*[1] article).Any such tool needs to be fit for purpose or, preferably, *best* for purpose.[2] The main statistical methods available for application in psychometric work have subtly differing objectives, and they are couched in terms of concepts such as reliability, validity and 'responsiveness to change'. In general the statistical methods (and designs) that are used in psychometric work are adaptations of the well-known standard statistical approaches, but with 'bespoke' labels reflecting their psychometric focus:

(1) Cronbach's alpha (α)
(2) Test-retest reliability ($r_{t\text{-}r}$)
(3) Item-total correlation ($r_{i\text{-}t}$).

When/why are psychometric statistical methods useful?

Psychometric statistical methods are useful first and foremost in optimising a new health scale that is being developed. However, once the iterative development process has resulted in an ultimate product, these methods should *also* be used to validate the scale that has been developed, to demonstrate formally to others that the scale is of good quality for research or clinical use. Such bench-marking, as fulfilling well the function intended (together with publication of a report of this validation exercise in an academic journal), will promote more widespread use of the scale by other researchers in that field. [Advantages will then accrue, for those wishing to use the health scale in their practice (both researchers and clinicians), if all research in a field tends to use the same good-quality tool. This is because there will be scope for amalgamation and interpretation of findings across separate studies, to provide more robust and dependable evidence.] Finally, these statistical methods are invaluable when an existing health scale, presumably already validated for the originally intended research/clinical context, needs adaptation/validation for use in new contexts (e.g. community use of a health scale developed for a hospital setting, or after translation of an existing questionnaire such as SF-36 into other languages).[3]

What precautions are needed?

It is common that a scale is developed initially by means of a respondent sample, who complete a 'development' scale comprising a pool of possible items (typically *more* items than are wished in the final scale). Iterative analysis is then undertaken to check the 'informativeness' and utility of each item, by applying something like item-total correlation. The least successful items would tend to be dropped at this stage.

As a general rule, once a (next) prototype scale has been decided, it needs to be validated (i.e. pertinent aspects of its validity and reliability need to be evaluated). At this stage there are likely to be further changes to the scale, and if these are more than minor then a further round of validation might be needed. It is not unusual for the various evaluations to lead to some conflicting suggestions, as to optimum action in respect of a particular item (e.g. if an item is relatively internally inconsistent with the rest of the scale,

Box 1 Glossary of statistical terms used in this article

Confidence interval (95% CI)	This defines a range of values within which we are 95% confident the true population effect lies.
Construct validity	See *Validity*.
Content validity	See *Validity*.
Correlation coefficient	See *Pearson correlation coefficient*.
Cronbach's alpha (α)	This evaluates *internal consistency*. It is calculated as the average of all possible split-half reliabilities, so it guards against any unlucky random choice that might occur if a single *split-half reliability* is calculated. In one sense it would seem that the higher the Cronbach α, the better (i.e. the more internally consistent the items). However, if α is too high, then one inference is that the items are *too* homogeneous, and there might well be some redundancy (and hence unnecessary burdening of future respondents with more scale items than needed). Therefore the recommended range for acceptable Cronbach α is between 0.7 and 0.9.
Face validity	See *Validity*.
Homogeneity of items	Tending to be scored in the same way, all fairly high, all middling, or all fairly low.
Internal reliability or consistency	See *Reliability*.
Item-total correlation ($r_{i((t-1)}$)	Used to quantify the homogeneity of responses to items in a scale that purports to measure a specific construct (e.g. anxiety). The standard *Pearson correlation coefficient* is calculated, in turn, for each item against the total score of the *remaining* items. The rule of thumb is that if any calculated r is <0.20 then that item would appear to be measuring something different to the rest of the scale, and consideration should be given to deleting it. However concern for *content validity* might influence retention of an item despite low item-total correlation.
Null hypothesis (NH)	A statement, prior to testing, of no effect.
Pearson correlation coefficient (r)	The standard parametric linear correlation coefficient (or product moment correlation) quantifying the strength of the relationship between two numeric variables (where r=1 is perfect correlation and r=0 no correlation).
Power	This term is used here, loosely, as the probability of rejecting the stated *null hypothesis* on the basis of the study data, when that is in fact the correct decision.
Precision	Accuracy of estimation possible from the study data (narrowness of *confidence interval*).
Reliability	Reliability encompasses a number of desirable features for a measurement scale. Most fundamental to these is minimal 'error' in responses. 'Error' occurs when an impression is given that differs from the truth, for example, if a respondent really does know something, or hold an opinion, but misunderstands the item, and erroneously responds negatively. Or if the respondent accidentally ticks the wrong response. Any process of measurement always runs a risk of some 'error', but scale design aims to minimise its probability (e.g. a layout that minimises inadvertently ticking the wrong response, wording that is as widely understood as is possible, and so on). Measurement error can be random (i.e. 'pure'), such as accidentally and randomly mis-ticking a response, or systematic (occurring *differentially*, that is, more in some respondents/circumstances, than in others), such as if less-educated respondents tend not to recognise a medical term for something they actually know. Both types of 'error' are of concern regarding measurement/diagnostic precision and research power, while the latter is a particular concern in that its occurrence will cause findings to deviate systematically from the truth (bias). Specific *types* of reliability are defined, including: *Internal reliability or consistency* – This is a measure of the homogeneity of items within a scale, the extent to which they are measuring a unitary construct. See also *Item-total correlation*. *Split-half reliability* – The items in a scale are randomly split into two halves and the Pearson correlation calculated between the total scores for each half. *Test-retest reliability* – This is a measure of the extent to which the same subject, if reassessed, will give similar responses on the two separate occasions, assuming the aspect being measured has not changed. The challenge is to judge the time interval between test and re-test, long enough to avoid the respondent recalling what they answered before (and reprising), but short enough to ensure that what is being measured has not changed in the meantime (e.g. in the interim the respondent might have watched a TV programme about ovarian cancer signs and symptoms).
Responsiveness to change	A formal process of evaluating whether the scale is responsive to an intervention that would be expected to alter the scores on the scale. Generally this is achieved by random allocation to two groups, with one group being 'primed' in some way (or 'treated') prior to completion of the scale – in the Simon *et al*.[1] study by being given a leaflet to read – and the scores of the two groups are then compared statistically against a *null hypothesis* of no effect being detectable.
Sensitivity to change	See *Responsiveness to change*.
Split-half reliability ($r_{s\text{-}h}$)	See *Reliability*.
Test-retest reliability ($r_{t\text{-}r}$)	See *Reliability*.
Validate	Show by purpose-designed research that in a specified context a scale/questionnaire that has been developed provides meaningful data about the health aspect(s) being 'measured', and that these 'measurements' are dependable/reliable. See also *Validity* and *Reliability*.
Validity	The characteristic that a scale does measure what it purports to measure. Specific *types* of validity are defined, some of the many variants being: *Construct validity* – The ability of the scale to measure an abstract concept (e.g. 'family complete') for which no absolute gold standard measure exists that could be used as reference standard. In such cases validity has to be evaluated indirectly, via a construct (e.g. if 'family complete' is true then irreversible contraception will be acceptable). For example, it might involve follow-up to ascertain percentages going on to choose irreversible methods of contraception, with construct validity indicated by greater percentages among those scoring high on the 'family complete' scale). *Content validity* – Expert judgement that the scale includes all aspects it should, given its aim, and does not include items addressing other (distinct) aspects. See also *Face validity*. *Face validity* – A special form of content validity, as assessed by experts.

but is needed for the sake of content validity). Therefore subtle judgement is often required, and this will need to be based on an understanding of the psychometric analyses, and an awareness of the research consequences of decisions that might be taken.

In the validation stage, it is preferable to use a *new* sample of respondents (not the development sample if there was one), and this sample should be representative of the population in whom the scale will be used. For example, if developing a diagnostic scale, validation should be on patients presenting as *possible* cases, who would in future be given the scale to complete, in order to examine whether the scale can differentiate true cases from *non-cases*. Regrettably it is all too often the case

that a validation sample comprises a group of true cases and a group of healthy 'controls'. In such a circumstance discrimination of cases would be trivially easy, and hence validation findings would be over-optimistic regarding the performance of the scale that could be expected in real clinical use. Not all scales are intended to be discriminatory, so in other scales different aspects of reliability and validity will be the focus of the validation research.

Example

In the Simon *et al*. study[1] what was to be assessed was 'public awareness of ovarian and cervical cancer'. An initial stage (and sample) to refine a pool of items was not required, since the health aspect being measured is knowledge about ovarian and cervical cancer, and the scale being developed is an adaptation of an existing generic Cancer Awareness Measure (CAM). The cancer-specific knowledge this scale should assess was decided from the literature, and reviewed by expert opinion (content validity). Item-total correlations were calculated and internal reliability was assessed by Cronbach's α. Test-retest reliability was assessed for those items for which this was technically possible (i.e. excluding open-ended items). Responsiveness to change was assessed by having a randomly selected subgroup of respondents read an information leaflet prior to completing the scale, and then comparing their scores with those for respondents *not* being given the leaflet. Construct validity was ascertained by a formal comparison of scores for 'standard' respondents against those for a group of 'expert' respondents selected on the basis that they were likely to be knowledgeable.

Overview

Validation of a health measurement scale is essential to ensure a useful and effective tool for execution of health research. The validation process involves fine-grained and detailed technical research and analysis.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

References

1 **Simon AE**, Wardle J, Grimmett C, *et al*. Ovarian and cervical cancer awareness: development of two validated measurement tools. *J Fam Plann Reprod Health Care* 2012;**38:167–174.**
2 **Streiner DL,** Norman GR. *Health Measurement Scales. A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press, 1995.
3 **Fukuhara S,** Bito S, Green J, *et al*. Translation, adaptation, and validation of the SF-36 Health Survey for use in Japan. *J Clin Epidemiol* 1998;**51**:1037–1044.