International Journal of Nursing Studies 47 (2010) 937–938

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

Commentary

Reliability assessment and approaches to determining agreement between measurements: Classic methods paper

Peter Griffiths *, Trevor Murrells

King's College London, National Nursing Research Unit, James Clerk Maxwell Building, Waterloo Rd, London SE1 8WA, United Kingdom

ARTICLE INFO

Keywords:
Reliability
Test–retest reliability
Inter-rater reliability
Measurement
Reproducibility of results
Validation studies
Observer variation

ABSTRACT

This classic methods paper (Bland and Altman, 2010) considers the assessment of agreement between measures, an often overlooked aspect of assessing measurements taken for use in research and practice and (re) introduces the ubiquitous 'Bland Altman' procedures for assessing agreement. The importance of these procedures is high and they address issues that are not always considered in research which uses measurement scales or describes the characteristics of scales developed for use in clinical practice. Many widely used approaches for reliability assessment can fail to consider the agreement between measures at all and can give an entirely misleading impression of an instrument's suitability for use in research or practice.

1. Introduction

In deploying a research measure it is necessary to establish that both the measurement instrument and procedures deployed in the study are sufficiently accurate and consistent for the purposes at hand. Where a novel method of clinical assessment is developed (for example a pressure ulcer risk assessment scale) it is vital to ensure that repeated measures of the same underlying condition, often undertaken by different clinicians, will agree sufficiently to make comparison between people and identify real change in an individual when it occurs. A number of questions emerge but the key issues can be summarised thus. If a measurement is applied repeatedly while the underlying state is static how much variation will there be in the result? If different operators (researchers or clinicians) or different approaches are used to take the measurement, do the results agree?

These are two related but distinct issues, although the two are often used interchangeably and are not clearly distinguished in much of the literature (Kottner, 2009; Kottner et al., 2009). Reliability refers to error (in the statistical sense) inherent in the scores. The same measurement taken on the same person repeatedly (or by different observers at the same time) should give the same value (provided that there is no underlying change) but there will inevitably be some degree of variation. This is the essence of reliability assessment. Commonly reliability is assessed using a measure of association such as a correlation coefficient. However, there is another aspect of variation between measures, which is the extent to which two measurements tend to *agree*.

At first sight the issue of agreement appears to be precisely the same issue as association and agreement is frequently classified as an aspect of reliability assessment. For example, the widely used texts by Polit and Beck refer to it as equivalence assessment and deal with it as an aspect of reliability (e.g., Polit and Beck, 2007). However, a correlation coefficient used as an assessment of reliability of an instrument fails to consider agreement at all and hence does not offer appropriate reassurance.

The issue is a fundamental one. What if different practitioners or methods give *systematically* different measurements? In this classic paper, first published more than 20 years ago, Bland and Altman clearly illustrate how

DOI of original article: 10.1016/j.ijnurstu.2009.10.001
* Corresponding author.
E-mail address: Peter.griffiths@kcl.ac.uk (P. Griffiths).

0020-7489/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijnurstu.2010.03.004

a correlation coefficient such as Pearson's *r* simply does not measure agreement because it is completely insensitive to changes in scale. In the extreme, two sets of measure can correlate perfectly and yet show complete disagreement. Imagine if one sphygmomanometer consistently measured blood pressure as 15 mmHg higher than another. This fact would certainly alert us to some error in the procedure or the equipment. However, in the absence of any random error, a study assessing the reliability of the measures would give a correlation coefficient of 1, we would conclude that the measure is reliable and be falsely reassured of the instrument's suitability for practice. The same issue would apply if an assessment scale, such as a measure of function or pressure ulcer risk, which was designed for clinical practice had items that were ambiguous and tended to be rated differently by different clinicians. This tendency would be masked, and possibly completely hidden, by a correlation coefficient which established reliability when real and clinically significant differences could occur simply because different people applied the scale.

Many key texts in nursing research give extensive coverage of correlation coefficients as measures of reliability but give relatively slight consideration to approaches to assessing agreement. Although under certain well defined conditions Pearson's *r* and inter-rater reliability coefficients can be similar, this contributes to a misleading impression that Pearson's *r* is a valid measure of agreement and may lead to continuing use. For some people it seems that the difference between the concepts of correlation and agreement has become blurred and any distinction between the two, which perhaps existed when the person was first introduced to the concepts, has been lost through the passage of time. For others the distinction was never made apparent. Therefore we reprint here in its entirety Bland and Altman's seminal paper describing their procedures for assessing agreement. The paper is a classic in many senses of the word—it has certainly been widely used and cited. The paper focuses on clinical measurement but the issues discussed and procedures described apply to most situations where reliability of scales is assessed and where such evidence is required. They certainly apply to the assessment of inter-rater reliability, which we should perhaps more properly take to calling reliability/agreement assessment, but also most situations where evidence of test–retest reliability/agreement is sought or required. The procedures described by Bland and Altman apply to any measures that can be treated as ratio or interval level and thus can also be usefully be applied to many long ordinal scales such as quality of life measures. I commend this paper to readers of the IJNS.

References

Bland, J.M., Altman, D.G., 2010. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1986 327 (8476) 307–310. Int. J. Nurs. Stud 47 (8), 931–936.

Kottner, J., 2009. Interrater reliability and the kappa statistic: a comment on Morris et al. (2008). International Journal of Nursing Studies 46 (1), 141–142.

Kottner, J., Dassen, T., Tannen, A., 2009. Inter- and intrarater reliability of the Waterlow pressure sore risk scale: a systematic review. International Journal of Nursing Studies 46 (3), 369–379.

Polit, D., Beck, C., 2007. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Lippincott Williams & Wilkins.