International Journal of Nursing Studies 50 (2013) 968–973

Contents lists available at SciVerse ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

On measurements and their quality: Paper 1: Reliability – History, issues and procedures

Jason W. Beckstead *

University of South Florida College of Nursing, USA

ARTICLE INFO

Article history:
Received 2 April 2013
Accepted 2 April 2013

Keywords:
Measurement theory
Reliability
Validity
Intervention research

ABSTRACT

This is the first in a short series of papers on measurement theory and practice with particular relevance to intervention research in nursing, midwifery, and healthcare. In this article I discuss reliability, its origins in classic measurement theory, important issues to consider when operationally defining reliability for a particular study, correlational procedures for assessing the reliability of data once collected, including test–retest reliability, split-half reliability, and Cronbach's coefficient α. Some important insights into reliability provided by attenuation theory are also offered.

1. Introduction

"When description gives way to measurement, calculation replaces debate". (Stevens, 1951, p. 1)

As the above quote from Stevens implies, determining and disseminating evidence-based best practices depends upon quantitatively demonstrating the efficacy and effectiveness of our interventions. This paper and the others in this series are intended to address this aim in general, and in particular by strengthening the reader's intuitive understanding of some key issues in measurement so that s/he can make informed, and hopefully, optimal choices when designing and conducting intervention research in nursing, midwifery and healthcare.

This aim is reflected in this paper and the others to follow by the emphasis placed on the foundations, rationale, and consequences of the decisions that the researcher makes. My objectives are to clarify why measurement quality is crucial to accurately assess the success (or failure) of interventions, and to demonstrate how the various decisions that researchers make regarding measurement (either explicitly or implicitly) can influence the quality of the data we obtain from intervention studies. Throughout this series I will assume that the reader has a passing familiarity with the fundamentals of statistical theory as they are applied in intervention research because this assumption is necessary in order to draw attention to the key issues in measurement and to avoid remedial digressions.

In this first paper in the series I discuss reliability, its origins in classic measurement theory, important issues to consider when operationalizing reliability in a particular study, procedures for assessing the reliability of data once collected, and how reliability can be considered when interpreting correlation coefficients.

The next section presents an overview of what has become known as classic measurement theory which leads into a discussion of measurement error and a definition of reliability. The algebra of covariance and correlation is reviewed in Section 3 as these concepts form the foundation of measurement theory. In Section 4 correlation-based procedures for quantifying reliability are described. Section 5 presents Cronbach's α in some detail and illustrates its workings with a brief example.

* Correspondence address: University of South Florida College of Nursing, 12901 Bruce B. Downs Boulevard, MDC 22, Tampa, FL 33612, USA. Tel.: +1 813 974 7667; fax: +1 813 974 5418.
E-mail address: jbeckste@health.usf.edu.

0020-7489/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijnurstu.2013.04.005

Attenuation theory is introduced in Section 6 to provide some important insights into reliability. Section 7 provides a brief summary and recommendations for further reading on alternative forms of reliability and recent developments in the statistical properties of α.

2. Classic measurement theory: a brief history and review

At is base, measurement is a procedure for identifying elements of the real world with the elements of an abstract logical system. Most often the abstract logical system in question is mathematics. Measurement theory is concerned with establishing a linkage between mathematics and the elements in the real world that we wish to study. Two terms encountered in all treatments of measurement theory are ''reliability'' and ''validity''. First, and foremost, it is crucial to realize that reliability and validity are properties of the data we obtain and how we use them, not properties of the devices used to obtain these data. In this series I focus on reliability. For a detailed discussion of validity in nursing research see Beckstead (2009).

2.1. In the beginning

One of the most fundamental tenets of measurement theory was first proposed by Carl Spearman in 1904 when working to develop a means to measure individual differences in intelligence, a stable person characteristic, or trait. Spearman conceived of every measurement, or observed score, x, as consisting of two components, a true score on the construct of interest, t, and an error score, e, so that $x = t + e$. Because they contain error, observed values of x are considered fallible. The true score, t, is the score that would be obtained under ideal or perfect conditions of measurement. Spearman's formulation has become known as true-score theory or classic measurement theory. Because both t and e are unknowns, this formula cannot be used to estimate the measurement error in the observed scores without further assumptions.

Assumption 1: for an individual, the construct being measured is constant (over some specified time period) and the errors in measurement are random. This suggests that if an individual were to be measured an infinite number of times a series of x values would result, each consisting of the same true score but differing due to different error scores. Being random, the expected value of the error scores is zero, $E(e) = 0$. *Assumption 2*: given assumption 1, the true score is equal to the expected value of the observed scores over an infinite number of repeated measurements (made under similar conditions), $t = E(x)$. *Assumption 3*: observed differences among individuals may be due to differences in their true scores or due to differences in their error scores. This implies that the variance of observed scores is a composite of the variance of true scores and the variance of error scores: $\text{Var}(x) = \text{Var}(t) + \text{Var}(e)$, and a little algebra shows that $\text{Var}(t)/\text{Var}(x) = 1 - [\text{Var}(e)/\text{Var}(x)]$. This last piece, the extent to which a set of measurements is free from random error variance is *reliability*. As a proportion, reliability can range from 0 to 1; it equals 1 when all the observed variance in a set of measurements is due only to true-score variance, that is, when there are no random errors of measurement, and it equals 0 when all the observed variance is due to random error variance. This definition implies that some measurement errors can be random while others are systematic. When measurement error is systematic, it is referred to as *bias*. In the remainder of this article we focus on random measurement error.

Measurement theorists have studied random measurement error for some time. In two seminal papers published in the same issue of the *British Journal of Psychology* in 1910, Carl Spearman and William Brown showed that when multiple measurements of a construct (e.g., intelligence, self-efficacy, emotional well-being, etc.) are obtained from the same individual and combined, the ratio of true-score variance to observed-score variance in the composite score increases. In other words, one way to reduce random measurement error is to create instruments that contain multiple measurements and to aggregate these measurements together when defining the observed score for an individual. Spearman and Brown also showed that as the number of similar measurements being combined increased, so does the reliability of the composite scores. In essence, this is why today most psychological scales are formed by aggregating responses to multiple items.

2.2. Classic forms of reliability

One critical decision to be made when designing intervention studies, and one that has implications for assessing reliability, is whether the key variables to be measured are trait or state variables. Traits, by definition, are stable person characteristics; they are not malleable. When the aim of an intervention is to change the dependent variable by experimental manipulation, it must therefore be considered as a state variable. Covariates, on the other hand, may be either trait or state variables.

The term ''reliability'' has been used over the years to refer to two distinct concepts in measurement theory, stability and equivalence. If we measure a sample of people twice on different occasions using the same instrument and calculate the correlation between the two sets of observed scores, the correlation coefficient may be interpreted as a *coefficient of stability*. Alternatively, the correlation between scores on two versions of the same instrument given virtually at the same time is a *coefficient of equivalence*, showing how closely the two versions yield similar measurements of the same trait within the sample of people. In both cases, reliability is quantified using correlation coefficients. Before illustrating procedures for doing so, our discussion will need to become a little more mathematical in nature. To strengthen our intuitive understanding we will need to consider the algebra of covariance and correlation.

3. On covariance and correlation

Variance is a summary statistic that indicates the variation in a set of scores from their mean (and standard