Journal of Research in Personality 69 (2017) 149–155

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

A practical guide to understanding reliability in studies of within-person variability

John B. Nezlek *

College of William & Mary, Williamsburg, USA
University of Social Sciences and Humanities, Poznań, Poland

ARTICLE INFO

Article history:
Received 14 October 2015
Revised 16 June 2016
Accepted 23 June 2016
Available online 25 June 2016

Keywords:
Reliability
Multilevel modeling
Diary studies

ABSTRACT

This article concerns how to estimate reliability (defined as the internal consistency of responses to a scale) in designs that are commonly used in studies of within-person variability. I present relevant issues, describe common errors, make recommendations for best practice, and discuss unresolved issues and future directions. I describe how to estimate the reliability of scales administered in studies in which observations are nested within persons, such as daily diary and "beeper" studies and studies of social interaction. Multilevel modeling analyses that include a measurement level can estimate the occasion-level (e.g., days or beeps or interactions) reliability of scales. In such models, items on a scale are nested within occasions of measurement and occasions of measurement are nested within persons.

1. Introduction

The carpenter's adage is "Measure twice, cut once", and psychologists have taken this advice to heart. Although there is some debate about exactly how many items one should use to measure a construct, with some exceptions (e.g., Robins, Hendin, & Trzesniewski, 2001), there appears to be broad agreement that constructs worth measuring are worth measuring with more than one item. Although desirable in terms of minimizing the variance associated with idiosyncratic characteristics of a single item, using multiple items to measure a construct raises questions about the extent to which the multiple items intended to measure a single construct in fact, measure a single construct.

Such judgments are usually made on the basis of some type of reliability analysis. Although reliability can be estimated in various ways, different methods have in common the goal of estimating the ratio of true to total variance. For scales that are perfectly reliable (1.0), all of the variance is true, none is error. In contrast, for scales that are totally unreliable (0.0), none of the variance is true, and by extension, scores on the scale are essentially meaningless.

This article discusses reliability in terms of the internal consistency of a set of items, a conceptualization along the lines of Cronbach's alpha, a widely used measure of the internal consistency of trait measures. The choice of internal consistency as a metric was dictated by various considerations. Within the context of studies of within-person variability, the consistency of scores across occasions (test-retest reliability) is probably not appropriate. First, the construct being measured is often assumed to be unstable – scores at t_1 are different from scores at t_2, and then at t_3, and so forth (e.g., Revelle & Condon, 2015). Second, there are the practical concerns associated with defining the "test" and the "retest" when someone provides data over multiple occasions. Alternate form reliability requires the use of different measures of the same construct, something that is typically not practical within studies of within-person variability.

In most studies of within-person variability data are collected from each participant on multiple occasions, and when deciding how to estimate reliability in such studies, it is important to take into account how such occasions are defined or selected. Broadly speaking, data can be collected on a fixed or random basis. If all participants provide data at the same time (defined either absolutely or relatively) and the specific times data are collected have some meaning, then this is fixed design, which is sometimes referred to as a crossed design. In contrast, if the specific times participants provide data are not important and have no meaning per se, this a random design, which is sometimes referred to as nested design.

A good example of a fixed design in the study of within-person variability is a longitudinal study. For example, a researcher is interested in the changes that occur between 3 and 6 years old and collects data from 100 children when they are 3, 4, 5, and 6 years old. In such a study, ages are not randomly sampled from

* Address: College of William & Mary, Department of Psychology, PO Box 8795, Williamsburg, VA 23187-8795, United States.
E-mail address: jbnezl@wm.edu

http://dx.doi.org/10.1016/j.jrp.2016.06.020
0092-6566/

the population of children's ages (perhaps 1–12), and the conclusions from the study are not meant to be generalized to other sets of four different years (e.g., 6, 7, 8, and 9 or 2, 4, 7 and 11). The target of inference is very specific.

Estimating reliability within the context of fixed designs has a long history and is relatively well-understood. Given this, and the fact that studies of within-person variability are increasingly using random (nested designs), and estimating reliability for random designs is not well-understood, most of this article concerns estimating reliability in nested designs. Estimating reliability within the context of fixed designs is discussed following consideration of estimating reliability within the context of random designs. At this point, it suffices to note that the structure provided by a fixed design offers opportunities that are not available for random designs.

In contrast to fixed designs, in random designs the specific times or dates data are collected are unimportant. For example, whether a daily diary study starts on one date or another typically has no meaning, and participants in a study can start on different days. If the dates have specific meanings, then this is a fixed design. Similarly, in "beeper" studies, following the work of Csikszentmihalyi and Larson (1987), the specific times people are prompted to provide data are not important. In fact, collecting data on a non-predictable basis probably reduce the impact of participants' expectations on their reports. An important consequence of random sampling of occasions is that, unlike the case for fixed design studies, occasions of measurement cannot be grouped or matched in random design studies, and I discuss some of the implications of this for estimating reliability in the next section.

The examples I have provided thus far concern what Wheeler and Reis (1991) called *interval-contingent* data collection. In such studies occasions of measurement are triggered by the passage of time (either at regular or random intervals). Wheeler and Reis also discussed what they called *event-contingent* data collection, studies in which occasions of measurement are triggered by the occurrence of a specific event. Social interaction diary studies in which social interactions trigger data collection (Wheeler & Nezlek, 1977) are good examples of this type of study. Individuals have different numbers of social interactions that occur at different times. Within-person variability can also be studied within the context of event-contingent designs, and the reliability of measures collected in such studies needs to be examined. Moreover, unless a researcher can provide a compelling case for some type of underlying fixed sampling strategy such studies need to be considered as random (nested) designs.

2. Estimating reliability when occasions of measurement are random (nested designs)

Many studies of within-person variability can be conceptualized as some type of nested design in which occasions of measurement (e.g., days of dairy) are treated as nested within persons. Multilevel modeling (MLM) is currently and broadly regarded as best practice for analyzing the data collected in such studies, and I focus on how to use MLM to estimate reliability within such studies. I focus on how to estimate reliability for what I will call the occasion-level of measurement. In an archetypal study in which days are nested within persons this would be the level-1 or day-level reliability. Estimating reliability for level-2 measures (i.e., person-level measures such as traits) is well-understood, and the methods of doing this do not need to be rehashed in this article. Later, I discuss how the basic principles I describe for estimating reliability for this example data structure can be applied to data structures that are organized in terms of different hierarchies, e.g., three-level designs.

I start by reviewing some of the different methods that have been used to estimate reliability in studies concerning within-person variability that used nested designs. I begin this review with a discussion of the fundamental flaws or weaknesses of some methods (rather than starting with recommendations for what to do) because discussing the shortcomings of what might seem to be perfectly reasonable methods will make it easier to understand the advantages of the methods I recommend.

In this review I do not cite specific studies in which the authors have used various flawed methods. I have no desire to criticize specific researchers, and given the widespread nature of the problems I describe, the selection of specific studies would be arbitrary and would draw inappropriate attention to those studies. Moreover, I do not summarize the existing literature to provide an indication of the relative use of different methods of estimating reliability within studies of within-person variability. Decisions about estimating reliability within this context should not be based on some type of survey of what researchers do. Such decisions should be based on the assumptions underlying different methods and on what is known about the accuracy of these methods.

For purposes of discussion, let's assume a four item measure of a construct collected once a day for two weeks. The construct, the number of items, and the number of measurement occasions are not relevant, but a general example will provide a context within which the issues can be discussed. The goal of the analyses is to estimate the occasion-level reliability of this measure. Controlling for between-person differences and for within-person differences in days, how well do these four items measure a single construct? We will put aside questions about validity.

2.1. Method 1: Estimating reliability based upon means calculated across all observations

In this method, a researcher calculates a mean for each of the four items (e.g., collapsed across the occasions each person has provided data), and uses these means to estimate the reliability of the scale perhaps by calculating a Cronbach's alpha. Although these analyses do provide a reliability estimate of some kind, this estimate does not describe the occasion-level consistency of responses to these four items. This estimate describes the consistency of mean responses across occasions, not the consistency of responses within occasions. If one is willing to assume that traits represent means of states, then such a reliability estimate would be similar to the reliability of trait level measure.

The fundamental flaw in this method in terms of estimating the occasion-level reliability (day-level, interaction-level, beeper-level, etc.) is that the reliability of the means is based on the between-person level relationships between mean scores for the four items. One of the foundational rationales for using MLM to analyze within-inperson relationships in studies such as this is that within-person relationships are mathematically independent of between-person relationships involving the same measures (e.g., Nezlek, 2001). Just as knowing the correlation between two measures at the between-person level tells us nothing about within-person level relationships between these same measures (and vice versa), knowing the between-person reliability tells us nothing about the within-person reliability. Relationships at different levels of analysis need to be modeled separately, and by extension, reliabilities at different levels of analysis need to be estimated separately.

2.2. Method 2: Estimating the reliability for each day of study and combining these estimates

In this method, a researcher treats each day of a study as a separate study, and estimates the reliability of set of responses for