Journal of Research in Personality 70 (2017) 45–55

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Full Length Article

What are other-rated scales composed of? Sources of measurement error and true trait variance in other-ratings of the Big Five

John F. Rauthmann

Department of Psychology, Wake Forest University, 415 Greene Hall, Winston-Salem, NC 27109, United States

ARTICLE INFO

Article history:
Received 14 September 2016
Accepted 12 May 2017
Available online 16 May 2017

Keywords:
Big Five
Reliability
Measurement error
Generalizability theory
Variance decomposition

ABSTRACT

Other-ratings of targets' traits may consist – besides true trait variance (TTV) – of different measurement error sources, particularly due to raters, scales, items, measurement times, and random fluctuations. Using Gnambs' (2015) and Ones, Wiernik, Wilmot, and Kostal's (2016) procedures for partitioning variance in scales due to measurement error, available meta-analytical data on Big Five other-ratings were analyzed. They showed relatively little TTV (0–13%), which was especially decreased by both low inter-rater reliability and convergent validity of Big Five measures. Accounting for both, TTV levels rose, but were still small to medium (4–26%). These findings provide important insights on what Big Five other-ratings are composed of and how such scale scores may be interpreted and treated in further analyses (e.g., trait-outcome relations).

1. Introduction

Personality psychology and assessment often rely on people's self-ratings about their traits (e.g., "I am an extraverted person"). While a self-rating approach is intuitive and economical, equating personality with people's self-ratings has been viewed critically (Funder, 1999; for a review, see Paulhus & Vazire, 2007). Indeed, different personality conceptualizations (e.g., Hogan, 1982, 1996; Hogan & Blickle, 2013; McAbee & Connelly, 2016; Roberts & Wood, 2006; Vazire, 2010; Wood & Roberts, 2006) distinguish between a person's self-concept or identity – as captured by self-ratings – and his or her reputation – as captured by other-ratings from other people (Connelly & Ones, 2010; Kenny, 1994). It has been shown that self-ratings and other-ratings capture different aspects of one's personality and perform differently when predicting behaviors and life outcomes (McAbee & Connelly, 2016; Spain, Eaton, & Funder, 2000; Vazire, 2010). One important aspect of personality scales, whether self- or other-rated, is their usefulness for diagnostic purposes or other substantive research interests. For example, we would want to use trait scale scores to predict other kinds of variables, such as behavior, well-being, health, life events, etc. A prerequisite is then that we are actually capturing a person's trait with those scales, not "noise" or measurement error (ME). Several frameworks exist to systematize the types of ME we can expect when measuring traits (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Ree & Carretta, 2006; Schmidt, 2010; Schmidt & Hunter, 2014; Schmidt, Le, & Ilies, 2003). ME types and the reliability estimates addressing them have already been meta-analytically examined for self-ratings (Gnambs, 2015). But what do other-rated scale scores actually capture? How much of the "real trait" of the rated target do they, on average, tap as opposed to different types of ME? This research examines, with the use previous meta-analytical data, the extent of ME due to different sources in other-ratings of the Big Five.

1.1. Background

Classical test theory maintains that variance in any observed score is an additive function of variance in "true" test scores and ME (Lord & Novick, 1968). This makes the concept of reliability, defined as the ratio of true score variance to total observed score variance or 1 minus the ratio of ME variance to total observed score variance, important for psychometrics and personality assessment. Put crudely, a reliable scale harbors less ME. Knowing the reliability of a scale is important because ME limits the usefulness of observed scores in several ways. First, if a scale captures predominantly ME, however defined, then the scale is measuring with low precision (and may possibly even be measuring something else). Second, ME constrains correlations between observed scale scores and other scores. In effect, the correlations are attenuated (i.e., not as high as they could be if there was no ME). Because most research is interested in true scale scores, corrections for ME can be done to uncover the relation between an observed score with another score if it were measured with perfect reliability (i.e., no ME).

E-mail address: jfrauthmann@gmail.com

http://dx.doi.org/10.1016/j.jrp.2017.05.002
0092-6566/

In light of these issues, understanding the extent and nature of ME is important. For example, Generalizability Theory (Cronbach et al., 1972) provides a framework to systematize different types, or facets, of ME. To illustrate potential ME types in other-ratings of targets' personality traits, the design in Fig. 1 serves as an example. As can be seen, several targets are rated by two raters on two scales, using three items each, at two time points. If we were to form trait scores from these data, the scores are composed of variance due to the actual individual differences in traits of different targets, on the one hand, and different forms of ME, on the other hand. Importantly, these forms of ME can be systematic (non-random) or unsystematic (truly random), and they are only considered ME because we are interested in true trait variance only. In this example, there are four sources of systematic ME due to fluctuations in raters, scales, items, and measurement time and one source of unsystematic ME due to random error.

1.2. Sources of measurement error and reliability coefficients in other-ratings

Table 1 summarizes the sources and indices of ME that may be present in other-ratings of traits. One form of ME is pervasive in any measurement and hence "accounted for" by all kinds of reliability coefficients: Random ME. *Random error variance* (REV) captures fluctuations in responses to the same item at a given measurement point (e.g., due to capricious cognitive, affective, or motivational states). Such fluctuations are unsystematic and hence usually not psychologically meaningful (especially as they are not supposed to be correlated with any of the raters' and targets' characteristics). On the other hand, the systematic ME types may offer some psychological insights.

One first and obvious ME source may stem from the raters (e.g., family, friends, acquaintances, strangers) performing the other-ratings of a target. Specifically, *rater-specific variance* (RSV) is not shared among raters and hence represents what is unique to a given rater's judgment. RSV affects the ratings of a single rater, but by using several raters, its effects can be cancelled out. Ideally, there would be perfect overlap, or consensus or inter-rater agreement, between ratings by different raters, but this is rarely if ever achieved (Connelly & Ones, 2010; Kenny, 1994). Thus, indices of *inter-rater reliability* (IRR), such as the commonly used intra-class correlation coefficient, are important quantifications of rater consensus.

Another source of ME can be the measurement time or occasion. Variations in ratings may occur as a function of different situations exerting an influence. Specifically, *transient error variance* (TEV) is not shared between measurement time points and hence represents what is unique to a given time point. By sampling ratings at several time points, TEV effects can be cancelled out. Ideally, there would be perfect stability between different measurement points, resulting in the same rating scores or rank-orders of targets.[1] Thus, *coefficients of stability* (CS), such as Pearson's correlation coefficient *r*, are important quantifications of scales' *retest reliability* when (perfect) trait stability across time is assumed.

Two other sources of ME have to do with the instruments used for measurement. One source concerns the items, the other the scale that is composed of those items. *Item-specific variance* (ISV) is not shared between different items, that is, an item obtains unique responses (possibly because of specific interpretations it may elicit) irrespective of the trait it is supposed to capture. Consequentially, using several items to form a scale is advisable because ISV effects can be cancelled out while systematic trait-related variance is supposed to accumulate. Ideally, all items would tap the underlying trait to the same extent and thus be perfectly correlated.[2] Thus, *coefficients of equivalence* (CE), such as indexes of internal consistency (e.g., α, ω) as well as split-half or parallel-test correlations, are important quantifications of the extent to which different items homogenize and may tap a common, latent trait.

In addition to ISV, *scale-specific variance* (SSV) may also contribute to ME. It occurs when two or more scales are not alternatives in tapping a common trait, but actually tap different traits. This can occur when jingle-fallacies have been committed conceptually (two actually different constructs obtain the same scale label) or when scales differ psychometrically (e.g., idiosyncratic focus on a narrower conceptualization of a trait domain and thus restriction of the item universe; different item formats and response scales; varying instructions). Ideally, different scales intended to measure the same trait would correlate perfectly (at least once accounting for the other ME types). Thus, *generalized coefficients of equivalence* (GCE), indexed by convergent correlations in multi-trait multi-method matrices, are important quantifications of to what extent scales measure the same trait.

1.3. Quantifications of ME

As can be seen in Table 1 and as explained above, there are different indices or coefficients of reliability depending on which ME type is accounted for (Le, Schmidt, & Putka, 2009; Schmidt et al., 2003). If variance in a scale score is "contaminated" by sources of ME such as variance due to raters, scales, items, time points, and random fluctuations, then the *true trait-related variance* (TTV) needs to be peeled out like layers of an onion. Different reliability coefficients represent different layers as they account for different ME types. Specifically, CE, CS, CES, GCE, and GCES are of interest to both self- and other-ratings.

CE concerns REV and ISV and is most often indexed by Cronbach's α (and sometimes by ω) capturing the homogeneity of items. CS concerns REV and TEV and is most often indexed by a retest correlation between the same scale administered at different time points. CES, the *coefficient of equivalence and stability*, combines CE and CS in that it accounts for REV, ISV, and TEV. GCE concerns REV, ISV, and SSV and is most often indexed by convergent validity correlations between alternative scales measuring the same trait. GCES, the *generalized coefficient of equivalence and stability*, goes another step further than the GCE and accounts for TEV in addition to REV, ISV, and SSV. When examining other-ratings, raters may introduce another source of ME. Thus, the GCES cannot be used as an index of TTV. Rather, RSV needs to be accounted for in addition to REV, ISV, SSV, and TEV. All of these coefficients have in common that they are estimated on a standardized scale from 0 to 1, like a correlation coefficient.

CE (.77–.85), CS (.80–.88), CES (.64–.77), GCE (.62–.76), and GCES (.49–.67) have already been meta-analytically examined for

[1] If such stability is quantified by Pearson correlation coefficients, then the absolute or actual value need not remain the same; it suffices if the rank-order of participants in the sample remain the same (despite possible normative developmental changes in an upward or downward fashion on the scale of a trait continuum). Additionally, the assumption that cross-time stability of a scale reflects its reliability only holds if we also assume that the trait being measured does not change (or changes normatively and uniformly across all participants in the sample in the same way, thus preserving rank-orders among participants). If participants actually change differentially on the trait and the scale picks up on this, a lower correlation between time points would be observed. This would not index a lack of reliability, but one of trait stability. Most often, the distinction of trait stability vs. scale reliability is ignored, especially because retest intervals are usually short or at least less than one year (while more profound trait changes are believed to take longer). This research also assumes that cross-temporal stability of a scale indexes its reliability, but mainly because the meta-analytical coefficients of stability used here were derived from studies with intervals less than one year.

[2] This holds only if we want to measure a purely unidimensional trait where breadth and heterogeneity of the trait domain sampled is not relevant. Rather, the most important criterion is the high intercorrelations among items which would additionally, and ideally, be Rash-homogenous.