Personality and Individual Differences 111 (2017) 96–105

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Item properties and the convergent validity of personality assessment: A peer rating study☆

Rachel A. Plouffe *, Sampo V. Paunonen †, Donald H. Saklofske

Department of Psychology, The University of Western Ontario, Canada

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form 25 January 2017
Accepted 28 January 2017
Available online xxxx

Keywords:
Test construction
Self-peer agreement
Convergent validity
Content saturation
Social desirability
Mean responses
Personality
Self-report

ABSTRACT

The current research evaluated the impact of personality questionnaire item content saturation, item social desirability, and mean item responses on the overall convergent validity of three well-known personality measures. Archival data representing groups of same-sex undergraduate roommate dyads were used for this research. Results demonstrated that content saturation, measured using item-total correlations, was the most consistent predictor of item convergent validity, measured using self-peer item response correlations. In order to predict outcome variables in education, clinical, and vocational contexts using scores on personality questionnaires, it is important for researchers to employ item selection procedures that take into account the item properties that affect the test's convergent validity.

1. Introduction

The most common method for assessing personality and individual differences for more than a century has been the self-report personality questionnaire (Jackson & Paunonen, 1980; Paunonen & Hong, 2015; Paunonen & O'Neill, 2010). Traditional methods of administering questionnaires include the completion of paper-based rating scales, in which participants indicate how representative a specific trait label, behaviour tendency, or attitude is to them (Holden & Troister, 2009). Modern technology allows for scales to be computerized and tailored to individual respondents, making them arguably one of the most efficient indicators of personality. Regardless of how they are administered, self-report questionnaires are an expedient way to assess an individual's attitudes and behaviours (Jackson & Paunonen, 1980; Paunonen & O'Neill, 2010).

The premise underlying self-report measures of personality is that individuals possess enough insight into their own psychological processes and past behaviours to make accurate judgments about their personality characteristics (John & Benet-Martinez, 2000; Paunonen & O'Neill, 2010). However, personality theory and assessment faced a paradigm crisis when a wealth of evidence introduced by critics of self-report personality testing revealed that individuals' responses to personality test items demonstrated little cross-situational consistency. That is, there appeared to be little stability of reported behaviour across time and situations (e.g., Mischel, 1968; Shrauger & Schoeneman, 1979). Furthermore, correlations between personality trait measures and relevant behaviours seldom surpassed a ceiling of 0.30 (Epstein, 1983; Jackson & Paunonen, 1985). The fundamental assumption underlying personality testing, which maintains that the characteristics and behaviours of individuals remain stable enough across diverse situations to classify them as enduring personality traits, was thus undermined (Epstein & O'Brien, 1985).

The "person-situation" debate provided the basis for a wealth of research concerning the improvement of traditional methods of personality test construction and assessment (Paunonen, 1984). Such improvements involve two fundamental psychometric requirements for sound personality measurement: the establishment of the measure's reliability and validity (Clark & Watson, 1995; Loevinger, 1957). Many of the apparent inconsistencies in personality test scores across time reported in some studies can be explained by the use of scales lacking in these psychometric properties (Jackson & Paunonen, 1985). One notable problem with past measures, for example, has been the use of self-report questionnaire items that tend to elicit socially desirable responding (Jackson, 1984).

Personality scales, even within the same omnibus questionnaire, typically do not have identical reliabilities or validities. The items on

☆ This research was partially supported by Social Sciences and Humanities Research Council of Canada Research Grant 410-98-1555 awarded to the second author.

* Corresponding author at: Department of Psychology, University of Western Ontario, London N6A 5C2, Canada.
E-mail address: rplouffe@uwo.ca (R.A. Plouffe).
† Deceased 29 December 2015.

http://dx.doi.org/10.1016/j.paid.2017.01.051
0191-8869/

two scales might look similar in style and format, be written by the same item writers, and be the result of the same statistical item selection strategy, yet the scales have different validities. One reason could be differential desirability in the scales' items, but there are other item properties that could be at work, including an item's difficulty, wording, direction of keying, face validity, content saturation, and more. The primary purpose of this study was to evaluate some of these psychometric item properties in terms of their contribution to the convergent validity of self-report measures of personality.

1.1. Evaluating the convergent validity of self-report personality measures

There are a number of ways to evaluate the validity of individuals' total scores on self-report personality inventories. The current study evaluates the convergent validity for a series of measures, which we do by comparing different methods for assessing the same construct and looking for agreement (Campbell & Fiske, 1959; Nunnally & Bernstein, 1994).

One way to compute convergent validity coefficients involves having an individual who is well acquainted with the target complete the same questionnaires in a peer rating format, and then to correlate the peer responses with the target responses to items (Holden & Troister, 2009; Paunonen, 1984; Paunonen & O'Neill, 2010). This well-acquainted individual could be a parent, friend, significant other, teacher, or sibling. High correlations between self- and peer ratings provide evidence for the measure's convergent validity (Campbell & Fiske, 1959; Foster & Cone, 1995). Using self-peer response convergence as a means to assess convergent validity has been successfully applied in a number of previous studies (e.g., Costa & McCrae, 1992; Funder, 1987; Funder & Colvin, 1988; Jackson, 1984; Paunonen, 1984; Paunonen & Kam, 2014). It is assumed that the peer rater will have been exposed to a number of behavioural cues about the target's personality, and that as these relevant cues increase, so too will self-peer agreement or convergence (Paunonen & O'Neill, 2010). This method of establishing convergent validity was employed for the purpose of the current research. Specifically, in the current study, convergent validity for a series of personality measures was established by correlating self-report responses to personality questionnaire items with roommate responses to the questionnaire items. Convergent validity, in this case, is defined as the self-peer correlations on individual personality questionnaire items. Higher self-peer convergence was indicative of higher convergent validity.

1.2. Item properties affecting convergent validity

Various psychometric item properties can have a demonstrable effect on the convergent validity of a personality questionnaire. These include, for example, item content saturation, item means, and item desirability. Many of the putative shortcomings of personality assessment described in the literature, such as the absence of findings of test-behaviour predictability, can be said to be due to a lack of consideration for these item characteristics (Jackson & Paunonen, 1980).

1.2.1. Content saturation

A goal in the construction of theoretical, construct-based measures of personality is to establish construct validity, defined as “the degree to which it measures some trait which really exists in some sense” (Loevinger, 1957; p. 685). More recently, Borsboom, Mellenbergh, and van Heerden (2004) have argued that a personality questionnaire can only possess validity if an attribute exists and if trait variation causally produces variation in test scores. An important consideration in establishing personality questionnaire validity is the extent to which items are content saturated. Content saturated items contain trait-relevant content, and the best ones are the most prototypical representations of their content domains (Paunonen, 1984). A general assumption in conventional personality scale construction is that single scales should measure single, unitary personality constructs. Thus, a highly content saturated scale will have high scale homogeneity, with all items representing trait-relevant content and not trait-irrelevant content.

One method for constructing personality questionnaires reflecting high content saturation involves employing factor analytic procedures in order to maximize item homogeneity and internal consistency (e.g., Briggs & Cheek, 1986; Paunonen, 1984; Paunonen, 1987). Items with the highest loadings on the largest factor underlying the scale's item intercorrelation matrix are inferred to be most content saturated. These items correlate more highly among themselves than do those with low loadings on the factor, which is likely due to their relation to a common theme – that is, the trait being measured (assuming irrelevant homogenizing factors such as desirability can be ruled out). Paunonen (1984) constructed ad hoc Personality Research Form (PRF; Jackson, 1984) subscales of varying length with varying content saturation by retaining items with high to low loadings on the first unrotated principal component extracted from the scale's items. Those scales constructed to reflect maximum content saturation (i.e., having the highest factor loadings) were more highly correlated with criteria such as peer ratings than were scales simply constructed to maximize items' contributions to the prediction of a relevant trait criterion.

Maximizing content saturation may also involve computing item-total correlations. Here, responses to individual items are correlated with total scale scores, and higher item-total correlations are then assumed to reflect more content saturation. This index is clearly linked to the above-mentioned factor loading index. Paunonen (1987) demonstrated that item loadings in a multiple group factor analysis, where each scale's items were assigned to their own factor, correlated in excess of 0.99 with item-total correlations.

Construct-based scale items that are most saturated with trait relevant content can be more valid than even criterion-based scale items, on which the development and selection of items is based primarily on how well they contribute to the prediction of a criterion variable (John & Benet-Martinez, 2000). Paunonen (1984) argued that such higher correlations between peer ratings and self-ratings on the more content saturated PRF items in his study were attributable to those items being most prototypical of the trait. In other words, content saturated items are highly salient and likely to be highly representative of concrete trait-relevant behaviours. In contrast, items low in content saturation might measure multidimensional content or ambiguous content, which might be difficult for respondents, be they selves or peers, to interpret consensually.

1.2.2. Item means

It is generally proposed that the optimal items to select in test construction are those with moderate means or *p*-values (popularity or probability of endorsement values). Items with moderate mean endorsement levels (e.g., around 0.50 on a binary True/False response scale, or 3 on a 5-point Likert scale) can demonstrate maximal observed score variance and respondent discrimination (i.e., how well the item distinguishes between respondents on the measured trait). On the other hand, items with extreme *p*-values (i.e., values close to 0.0 or 1.0, or to 1 or 5 on the 5-point scale) fail to differentiate between individuals because of the restricted variance of item responses. Furthermore, items with extreme *p*-values impose limits on the strength of the correlations that the measure can demonstrate with criterion variables, thus attenuating indices of validity (Epstein, 1983).

Holden, Fekken, and Jackson (1985) examined the relationship between absolute endorsement frequency of 80 binary PRF items and criterion validity. Their results demonstrated a significant correlation of −0.29 between extreme endorsement levels and criterion validity. Thus, items that are endorsed by many respondents or by few respondents hinder the criterion validity of a measure (Nunnally, 1978). This does not mean that items with moderate means are definitely more valid, but rather such items do not have the same statistical constraint on validity.