Using principal component scores reduces the effect of socially desirable responding

Kalev Saar [a], Toivo Aavik [a,d], Kenn Konstabel [a,b,c,d,*]

[a] Institute of Psychology, University of Tartu, Tiigi 78, Tartu 50410, Estonia
[b] National Institute for Health Development, Hiiu 42, 11619 Tallinn, Estonia
[c] College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[d] Estonian Centre of Applied Psychology, Tartu, Estonia

ARTICLE INFO

ABSTRACT

Konstabel, Aavik, and Allik (2006) have found that controlling for social desirability in self- and peer-reports of personality using NEO Personality Inventory results in higher consensus. We report a reanalysis of these data showing that a similar effect is achieved when factor scores are used instead of unit-weight scores. The factor scores were also closer to being orthogonal even though they were computed using the coefficients published in the questionnaire's manual. These findings are interpreted in terms of a general evaluative bias that is more or less unique to each rater, as opposed to trait information that is largely consensual.

1. Introduction

The validity of questionnaire data has always been an important concern in research on personality. Self-report questionnaires are potentially fallible instruments—respondents may answer at random, they may misunderstand items, agree or disagree hastily, or fake deliberately. There is plenty of evidence of the validity and usefulness of self-report personality trait scores (e.g., Connelly & Ones, 2010; Ones, Dilchert, Viswesvaran, & Judge, 2007; Paunonen & O'Neill, 2010) but also of their bias (e.g., Anusic, Schimmack, Pincus, & Lockwood, 2009) and fake ability (e.g., Ferrando & Anguiano-Carrasco, 2011). Several algorithms have been proposed to make better use of the questionnaire data; a number of them are specifically targeted at diminishing the influence of socially desirable responding. Some of them have concentrated on "removing" biased information: partialling out social desirability of items before computing scale scores (Saucier, 1994), wiping away the first general factor (Paulhus, 1981); others have suggested strategies of revising or selecting items (Bäckström, Björklund, & Larsson, 2009; Petterson & Turkheimer, 2010). Removing a large portion of variance from personality scales may reduce their validity; the significance of this concern needs to be explored using some external criterion, for example, acquaintance reports ("consensual validation", McCrae, 1982).

There is a common understanding that peer-reports are among the most relevant validity criteria for self-reported traits. Self-report data typically show considerable correlations with reports provided by acquaintances (McCrae & Costa, 2003), and the agreement tends to improve as informants have more contact with the target person (Funder & Colvin, 1997; Kenny, 1994; Paulhus & Bruce, 1992). Thus, reports by well-acquainted peers could be used as one criterion for the validity of self-report data. Peer-reports, however, are not invulnerable of social desirability: Konstabel et al. (2006) showed that partialling out social desirability (operationalized as sum of items weighted by their social desirability values, as proposed by Hofstee, 2003) from both self- and peer-report data improves self-peer as well as peer–peer correlations.

In a different context, Costa and McCrae (1992) have advised using factor scores instead of unit-weight scores, for two reasons: (1) "factor scores are more nearly orthogonal"; (2) "... tend to have somewhat higher validities against external criteria" (McCrae & Costa, 1989). But why should "more nearly orthogonal" scores be desirable? A purely methodological reason is that the closer the scores are to orthogonality, the larger is the share of the original variance they summarize; therefore, orthogonal scores are more economical. A second possible reason is that in case of unit-weight scoring, the common variance (be it a "true" effect or a shared bias) is repeated in all scores, whereas it is divided between components if orthogonal factor or principal component scores are used. One might thus expect that using factor scores might have an analogous improving effect to controlling for social desirability index.

A rigorous proof of this intuitive argument would presume a sound psychometric model of how evaluative bias operates in

* Corresponding author at: Institute of Psychology, University of Tartu, Tiigi 78, Tartu 50410, Estonia. Tel.: +372 7375 903; fax: +372 7376 152.
E-mail address: kenn.konstabel@ut.ee (K. Konstabel).

self- and peer-reports, and a systematic exposition of how, and under which conditions the effect appears. For the present purpose, however, we just need to show that the above reasoning is not pure speculation and that under some thinkable conditions, using principal component scores instead of unit-weight scores is expected to reduce the general evaluative component (provided that such variance component is there in the data). We therefore ran a simulation with four "items" (A1, A2, B1, and B2) reflecting two "traits" (a and b), a common "evaluative bias" (sd), and "random error" ($e1$ to $e4$). The "traits", "bias", and "error" were generated at random, whereas the "items" were just unweighted sums of three components: the "trait" (a in A1 and A2, b in B1 and B2), an "error" component unique to each item, and finally, the "evaluative" component common to each item. A trivial expectation is that in a principal components analysis, each of the two rotated components would reflect a mixture of one of the traits (a or b) and the general evaluation. In the simulation we compared the correlations between unit-weight scores (A = A1 + A2, and B = B1 + B2) on the one hand, and two Varimax-rotated principal component scores on the other hand, with the general "evaluative bias"—that is, the random variable used in generating the items. With the above described setup, the correlation between unit-weight scores A and B was, on the average, 0.4, whereas principal component scores were obviously uncorrelated. The mean correlation between unit-weight scores and the "evaluative bias" was 0.63, and the mean correlation between principal component scores and the "evaluative bias" was 0.53; in 1000 simulations, the former correlation was always bigger. The corresponding unit-weight scores and principal component scores were highly intercorrelated: on the average, 0.97. R code to produce this simulation is given in the web appendix.

While the above simulation is based on conveniently selected assumptions rather than a sound psychometric model, it does show that under certain conditions, using PC scores instead of unit-weight scores is expected to reduce a general evaluative bias. The purpose of the present paper is to test this idea empirically. As a methodological note, we would not expect large differences between different factor analytic methods and principal components analysis (PCA), but will use PCA in the present paper as the computation of PCA scores is simpler and does not require an iterative algorithm.

In the present paper we reanalyze data from Konstabel et al. (2006) to explore (1) whether using principal component scores instead of unit-weight domain scores improves self-peer and peer–peer correlations, and (2) whether controlling for social desirability index (Hofstee, 2003; Konstabel et al., 2006) would improve consensus even if principal component scores are used. It is also expected that using principal component scores instead of domain scores would reduce intercorrelation between the five factors.

2. Method

The present analysis is based on data from Konstabel et al. (2006). Three hundred and ninety students from University of Tartu, Estonian Agricultural University and Mainor Business School participated in the original study (318 women and 72 men; mean age 22.3 years, SD = 5.2). Some of the students received partial course credit for their participation.

Respondents were assigned to one of two conditions and asked to complete the Estonian version (Kallasmaa, Allik, Realo, & McCrae, 2000) of the NEO-PI-R (Costa & McCrae, 1992) measuring the Big Five dimensions of Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. In the applicant condition ($N = 164$), participants were asked to respond as if they were applying for a job, and their questionnaire contained the following instruction:

You have all the skills to perform well in this job, but there are many other candidates whose skills are equal to yours and who are highly motivated to get the job. Please answer the following questionnaire in the way you would do in this situation.

In the honest condition ($N = 226$), participants' questionnaires contained an instruction to describe themselves honestly and accurately.

All participants were asked to recruit two acquaintances to complete peer-report versions of NEO-PI-R about them. The peer-report instructions contained no reference to the job application situation. At least one peer-report was available for 376 participants and two peer reports for 289 participants. If two peer-reports were available, the average of 2 peers was used for computing self-peer consensus correlations.

Social desirability ratings were independently collected from 88 judges (24 men and 64 women, mean age 37.6 years), who assessed the desirability of each of the 240 NEO-PI-R items on a 7-point Likert scale. These ratings were used to compute the social desirability index proposed by Hofstee (2003). For this purpose the social desirability ratings and participants' responses to each item were first linearly transformed to a scale from −1 to +1, and then multiplied by each other. Finally, each participant's social desirability indices (SDI) were averaged. In the resulting score, socially desirable responding is conceptualized as the degree to which a person agrees with socially desirable items and disagrees with undesirable items.

Factor score coefficients from the NEO-PI-R manual (Costa & McCrae, 1992) were used to compute the factor T-scores. We used published factor score coefficients instead of ones derived from the present study because of practical considerations: in applied uses of personality scales, one would seldom have groups that are large enough for factor analysis and testing of single individuals is common.[1]

All data analyses were carried out with statistical software R version 2.12.2 (R Development Core team 2010), using add-on psych (Revelle, 2010) and boot (Canty and Ripley, 2010).

3. Results

We first tested (Table 1) whether using principal component scores has an effect on correlations of NEO PI-R domain scales with three indices of socially desirable responding: the social desirability index (SDI, proposed by Hofstee, 2003), and two subscales of the Balanced Inventory of Desirable Responding (BIDR: Paulhus, 1991): self-deceptive enhancement (SDE) and impression management (IM). In most cases, these correlations were significantly lower in absolute value when principal component scoring was used; especially notable was the reduction in correlations of SDI with Neuroticism and Conscientiousness. A reverse effect occurred in two cases (Agreeableness with SDE, and Extraversion with IM) but is difficult to interpret.

In Table 2, we present the effects of faking instruction on the means of NEO PI-R domain scores. The effects were generally smaller when principal component scoring was used, except for Openness which was not affected by faking instruction.

Correlations between principal component scores were generally higher than corresponding correlations between unit-weight scores (Table 3). The only exception was Extraversion in peer–peer agreement. The $\bar{Z}_2^2$ statistic (Steiger, 1980) was used to test the significance of differences between individual zero-order and principal-component score correlations. In both honest and applicant conditions three out of five traits showed significant improve-

---

[1] The results did, however, lead to the same conclusions when the principal component score coefficients derived from the present sample were used (data presented in the web appendix).