Journal of Research in Personality 53 (2014) 54–63

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Brief Report

Selecting items for Big Five questionnaires: At what sample size do factor loadings stabilize?

Gerrit Hirschfeld [a,b,*], Ruth von Brachel [c], Meinald Thielsch [d]

[a] Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrück, Osnabrück, Germany
[b] German Paediatric Pain Centre, Children's Hospital Datteln, Germany
[c] Department of Clinical Psychology and Psychotherapy, Ruhr-University Bochum, Bochum, Germany
[d] Department of Psychology, Westfälische Wilhelms-Universität Münster, Münster, Germany

ARTICLE INFO

Article history:
Available online 27 August 2014

Keywords:
Personality assessment
Big Five
Exploratory factor analysis
Sample size
Sequential sampling
Bootstrapping
Simulation study

ABSTRACT

Researchers often use exploratory factor analysis (EFA) to develop and refine questionnaires assessing the Big Five personality traits. We use sequential sampling and bootstrapping to determine the number of participants needed to yield stable loading patterns for the Big Five Inventory (BFI) and the International Personality Item Pool Big Five measure (IPIP). Overall 21,350 participants (BFI = 10,285; IPIP = 11,065) participated. In two studies primary factor loadings are highly variable in smaller samples ($n < 500$) and some primary loadings are not stable with 10,000 participants. Most studies will not have adequate sample size to yield stable loading patterns for Big Five measures such as the BFI and IPIP. Researchers should assess and report the variability of loading patterns.

1. Introduction

Assessment of personality is largely based on the five factor model (cf. John & Srivastava, 1999). Items for questionnaires that aim to assess these factors are often identified using factor loadings from an exploratory factor analysis (EFA) as criteria. Similar strategies are used when shorter versions are developed. For example Rammstedt and John (2007) selected items that among other criteria exhibited a simple-loading pattern, i.e. items showed substantial loadings on only one factor and are no substantial cross-loadings to other factors. While these criteria for item selection make intuitive sense, several studies reporting item-level analyses of the NEO-Five Factor Inventory revealed that several items included existing inventories do not meet the simple-structure criterion (Egan, Deary, & Austin, 2000; Parker & Stumpf, 1998) and argued for a removal of these items. Even though theses studies used large samples ($n > 1000$) chance variability of item loadings may tempt researchers to develop different questionnaire versions based on empirical findings. Hence the stability of factor loadings is an important aspect, since it may mislead researchers into developing concurrent versions of existing measures.

Of course many previous simulation studies have dealt with the question of how many participants need to be analyzed to yield stable factor loadings (for a review see: Beavers et al., 2013; Guadagnoli & Velicer, 1988; MacCallum, Widaman, Zhang, & Hong, 1999). However, it is hard to interpret these findings, if the goal of the researcher is to decide which items to include in an instrument, since the indices to describe agreement between loading patterns such as *g* or kappa are aggregates over all item loadings. In contrast to this, researchers who are using EFA to develop novel or enhance existing inventories need to know whether and from what sample size on traditional decisions rules, e.g. loadings >.3 and no cross loadings <.3 yield stable recommendations about individual items.

The aim of the present study is to estimate how many participants are needed to achieve stable loading patterns, e.g. loadings patterns that do not change any more due to the inclusion of additional participants, in two datasets that are typical for personality psychology. We take a novel sequential sampling approach and a more traditional bootstrapping approach to address this question. The sequential sampling approach is inspired by recent work on the stability of correlations (Schönbrodt & Perugini, 2013). In their simulation study the authors estimated the correlations coefficient between two variables for a growing number of participants and inspected the trajectory, i.e. development with growing sample size, of

* Corresponding author. Address: Faculty of Business Management and Social Sciences, University of Applied Sciences, Caprivistr. 30, 49076 Osnabrück, Germany.
E-mail address: g.hirschfeld@hs-osnabrueck.de (G. Hirschfeld).

http://dx.doi.org/10.1016/j.jrp.2014.08.003
0092-6566/

these estimates. For each trajectory they determined a point of stability as the sample size from which on the parameter estimates converge on a true value. Using typical effect size estimates and definitions of what constitutes negligible deviations form the true value, they concluded that 250 participants are needed to estimate correlations with some confidence. Since EFAs build on correlation matrices, this result is also relevant to sample size planning for factor analysis. Extending this methodology, we first describe the trajectory of factor loadings from participants' responses. Second, we implement the simple structure criterion to determine the point of stability (POS), i.e. the sample size from which on researchers can decide with certainty whether to drop or retain an item. Third, we generate different random orders to assess the variability of POS. Fourth, we compare the results of this analysis to more traditional bootstrapping analysis in which pseudosamples of different sizes are drawn (with replacement) and the variability of factor loadings is inspected. We believe that the results of these analyses are directly relevant to researchers in personality psychology.

2. Study 1 – Big Five Inventory

The first study investigates the stability of factor loadings using a German version of the Big Five Inventory (BFI: John, Donahue, & Kentle, 1991).

2.1. Methods

2.1.1. Subjects

The subjects were recruited via a German online panel for psychological research (called PsyWeb, available through https://psyweb.uni-muenster.de). Overall 10,285 participants completed the BFI and were included in the analysis. Participants were between 14 and 85 years old (39 ± 14). About 53% (n = 5225) of the participants were female, reflecting the trend in German demographics.

2.1.2. Materials and procedure

Participants were invited via e-mail to an online survey. After a welcome and an instruction page the BFI was presented. We used a slightly adapted version of the German BFI version of Lang, Lüdtke, and Asendorpf (2001). The version of Lang and colleagues consists of 42 items representing the five factors (E = Extraversion; C = Conscientiousness; N = Neuroticism; O = Openness to experience; A = Agreeableness). We added two items ("I am someone who can be moody" and "I am someone who likes to cooperate with others") from the original BFI version, which had 44 items. Additionally, we rephrased another four Items ("I am someone who is enthusiastic and inspires others", "I am someone who works reliable and conscientious", "I am someone who is emotionally stable, and not easily upset", and "I am someone who remains calm in stressful situations") to better grasp the meaning of the original BFI. All items were presented in a random order for each participant. At the end of the survey participants were asked for some demographic data and for their permission to use the data for scientific purposes. Independently from the answer in this permission all participants received an automatic but personalized feedback on the Big Five dimensions. Completing the whole survey (including feedback) took the participants on average about 8 min.

2.1.3. Data analysis

The data were analyzed in four steps. First, the factor solution forcing the extraction of five factors and varimax rotation for the whole sample was computed using the *factanal* function in R. This solution was used in later steps as the standard against which intermediate solutions were compared. Second, we determined a "trajectory" of factor loadings in a specific sample by repeating the analysis for different subsets of participants. Specifically, we sequentially added participants one by one (from 50 to 10,285) to the dataset and computed the factor loadings for each sub-sample, i.e. the first 50 participants then the first 51 participants until all participants were added. The last analysis in this trajectory corresponds to the analysis of the whole sample. Instead of plotting the raw factor loadings, factor loadings for items with primary loadings smaller than .3 or cross-loadings larger than .3 were assigned a loading of zero. This was done in order to implement a decision rule, according to which only items that conform to simple structure should be retained. Even though there is no objective definition for substantial loadings several guidelines suggest that only items should be retained that have primary loading of at least .3 on the target factor and no cross loadings larger than about .3 (e.g., Nunnally, 1978; Tabachnick, Fidell, & Osterlind, 2001). Based on this rule we calculated the point of stability (POS) as the sample size from which on the loading pattern was stable. For items that showed a primary loading larger than .3 in the final sample, the POS was the sample size from which on all following analyses yielded primary loadings larger than .3. For items that showed primary loadings smaller than .3 or cross loadings larger than .3 in the final sample, the POS was the sample size from which on all following analysis yielded smaller primary loadings or cross-loadings. Since all participants were added exactly once, the last estimates for the primary loadings were identical to the estimates in the whole sample. Because of this all items in all orders had a POS that was smaller than the maximal sample size. Third, we generated 1000 different orders of participants for which we determined trajectories of primary loadings and calculated the associated POS. This was necessary to account for the fact that the POS is sample specific, i.e. extreme participants may be sampled early or late within a specific order and thus distort the POS. From the distribution of POS values across the different orders we estimated the 10%, 50%, and 90% percentiles as, optimistic, average and conservative estimates of the POS. Fourth we performed a bootstrapping analysis to assess the variability of the factor loadings at selected sample sizes (50, 100, 250, 500, 1000, 1750, 2500, 5000, 7500, 10,285). In this analysis we drew (with replacement) 5000 pseudosamples with a specific sample size. Since this may include several participants twice or more often, this technique is suited to assess the variability at the full sample size. For each pseudosample the same simple-structure criterion was used to determine in how many of the pseudosamples an item would have been retained. We deemed the primary loading to be stable at a specific sample size if the probability of being retained was larger than 90% or smaller than 10%. That is if the item was retained in more than 90% of the samples or dismissed in more than 90% of the samples.

All in all 10,285,000 (10,235 subsamples for each trajectory * 1000 random orders + 50,000 bootstrapping samples) EFAs were computed for the first BFI-dataset. All analysis were performed in R (R Core Team, 2012), and the code can be found as an appendix.

2.2. Results

2.2.1. Factor solution in the whole sample

Overall all items loaded substantially on the proposed factors (Table 1). However, eight of the 44 items also showed cross-loadings larger than .3, that would lead some researchers to drop these items for a final version of the questionnaire. Furthermore, especially the items for the agreeableness factor showed loadings that were only merely above the threshold.