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

Gerrit Hirschfeld[1], Ruth von Brachel[2], Meinald T. Thielsch[3]

[1] German Paediatric Pain Centre, Children's Hospital Datteln, Germany
[2] Ruhr-University Bochum, Germany;
[3] University of Münster, Germany

Highlights:

- EFA is widely used to develop and refine Big Five inventories
- We investigate the stability of loading patterns with sequential sampling and bootstrapping
- Some items yield unstable loading patterns with fewer than 1,000 participants
- Researchers need to take into account variability of the factor loadings

Corresponding author:
Dr. rer. nat. Gerrit Hirschfeld; German Paediatric Pain Centre, Children's Hospital Datteln, Dr.-Friedrich-Steiner Str. 5, 45711 Datteln, Germany; Tel.+49- 2363-975-183; Fax.+49- 2363-975-181; eMail: g.hirschfeld@deutsches-kinderschmerzzentrum.de

Please cite as:

Hirschfeld, G., von Brachel, R. & Thielsch, M. T. (2014). Selecting items for Big Five questionnaires: At what sample size do factor loadings stabilize? *Journal of Research in Personality, 53,* 54-63. doi: 10.1016/j.jrp.2014.08.003

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.

Keywords: personality assessment; big five; exploratory factor analysis; sample size; sequential sampling; bootstrapping; simulation study

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 Rammsted 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 > 1,000) 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 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 & 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, Lütke 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 eight minutes.

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 1,000 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, 10285). In this analysis we drew (with replacement) 5.000 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 pseudo-sample 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 * 1,000 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 (tab. 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.

2.2.2 *Trajectory for one specific sample*

The trajectory for the loadings for one specific sample is given in figure 1. As indicated by the vertical lines, several cross-loadings larger than .3 occur in smaller sample sizes. The loading patterns are relatively unstable in sample sizes smaller than 1.000 participants. With larger sample sizes the loading pattern is stable for all items from the conscientiousness, neuroticism, and openness for experience factors. Individual items from the agreeableness and extraversion factors do not show a stable pattern in sample sizes of about 7.000 participants.

2.2.3 *Point of stability in 1,000 random orders*

In order to test the stability of these findings, we generated 1,000 random orders in which the participants were sequentially added to the analysis. The distribution of the POS across the different orders for the individual items is depicted in figure 2 and quantiles are given in table 1. Specifically, 6 items had a median POS larger than 500. However, there was considerable variability in the POS depending on the order in which the participants were added as indicated by the distance between the percentiles. For example E3 shows a stable loading pattern from the beginning in the "best" 10% of the orders, while at the same time needing more than 2000 participants to stabilize in the "worst" 10% of the orders. Critically, the 90% percentile as a more conservative estimate of the POS indicated that 9 items needed more than 1,000 participants to yield stable estimates.

2.2.4 *Bootstrapping results*

In order to assess the variability of factor loadings at specific sample sizes, a bootstrapping analysis was performed. The proportion of pseudosamples in which an individual item was found to have a stable primary loading is depicted in figure 3. Overall the proportion of pseudosamples converged to either 1 for items with simple primary loadings or 0 for items with small primary loadings or substantial cross loadings. With 500 participants, ten items had unstable primary loadings, in the sense that they were retained or dismissed in more than 90% of the samples. With 1,000 participants, eight items had unstable primary loadings. Three items - two from the agreeableness factor and one from the extraversion factor - had unstable primary loadings using the full data set.

Table 1. Factor loadings for full sample and points of stability.

	Factor loadings					Point of stability		
						10%	median[c]	90%
Item	Factor1	Factor2	Factor3	Factor4	Factor5	percentile		percentile
E1	.716	.004	.156	.01	-.154	50	60	200
E2	-.788	.063	-.054	-.034	.106	50	50	91
E3	.495	-.219	.26	.333	-.002	50	376	2007.5
E4[b]	.578	-.083	.375	.111	-.015	50	103.5	448
E5	-.795	.008	-.079	-.043	.096	50	50	86.1
E6	.508	-.224	.163	.296	.291	3165	7798	9528
E7	-.613	.253	-.005	-.153	.021	50	21.5	1025.5
E8	.75	-.072	.118	.014	-.156	50	51	170
C1	-.053	.019	.045	.73	-.072	50	60	200
C2	-.003	.052	0	-.414	.333	50	50	91
C3[b]	-.027	.022	.024	.727	-.107	50	50	86
C4	-.038	.019	.129	-.513	.079	50	531	2517.8
C5	-.198	.077	-.02	-.554	.144	50	50	101.1
C6	.065	-.133	.148	.596	.016	50	56	128
C7	.169	-.081	.055	.604	-.003	50	89.5	264.2
C8	.249	-.18	.156	.494	.044	50	59	158
C9	-.04	.238	.042	-.512	.044	50	58	152
N1	-.365	.512	.037	-.197	.13	63	232	994
N2	.044	-.77	.074	.077	-.048	50	135	641.3
N3	-.115	.638	-.019	-.047	.279	50	50	86
N4	-.217	.591	.054	-.028	.031	50	92.5	566
N5[b]	-.006	-.749	.049	.089	-.184	50	50	96
N6[a]	-.063	.479	.028	-.141	.41	50	767.5	3816.3
N7[b]	-.027	-.698	.11	.1	-.078	50	88	40.1
N8	-.362	.596	-.032	-.2	-.008	50	50	169.3
O1	.226	-.154	.729	.066	.082	50	105.5	339.6
O10	.038	.031	.434	-.036	-.124	50	50	117.1
O2	.188	-.135	.466	.093	-.08	50	101	56.1
O3	-.14	.135	.39	.048	-.022	50	92.5	566
O4	.169	.018	.658	-.044	-.026	50	50	96
O5	.161	-.139	.711	.084	.081	50	113	448.6
O6	.017	.054	.56	.001	-.151	50	99	30.1
O7	-.187	.259	-.307	-.157	-.044	50	102.5	246
O8	.079	-.115	.601	.014	.072	50	132	431.3
O9	-.041	-.043	-.547	.03	.176	50	68	165.1
A1	.08	.196	.042	-.021	.503	50	74	187.1
A2	.062	.047	.17	.095	-.35	50	77	24.2
A3	.125	.292	.03	-.093	.466	557.5	5057	8832.5
A4	.117	-.19	.058	-.043	-.318	50	58	130
A5	.289	-.126	.078	.033	-.339	50	89	308.3
A6	-.282	.022	.009	-.057	.577	50	114	395.1
A7	.037	.061	.187	.119	-.51	65	368.5	1791.1
A8	-.132	.138	-.008	-.056	.694	292.9	4534.5	8445.7
A9[a]	.373	-.121	.089	.116	-.302	125.7	1377.5	5397.8

Note: E = Extraversion; C = Conscientiousness; N = Neuroticism; O = Openness to experience; A = Agreeableness; [a] = new Item in German version, based on the original BFI; [b] = changed Item in German version to better reflect the original BFI; [c] = fractions indicate ties in data

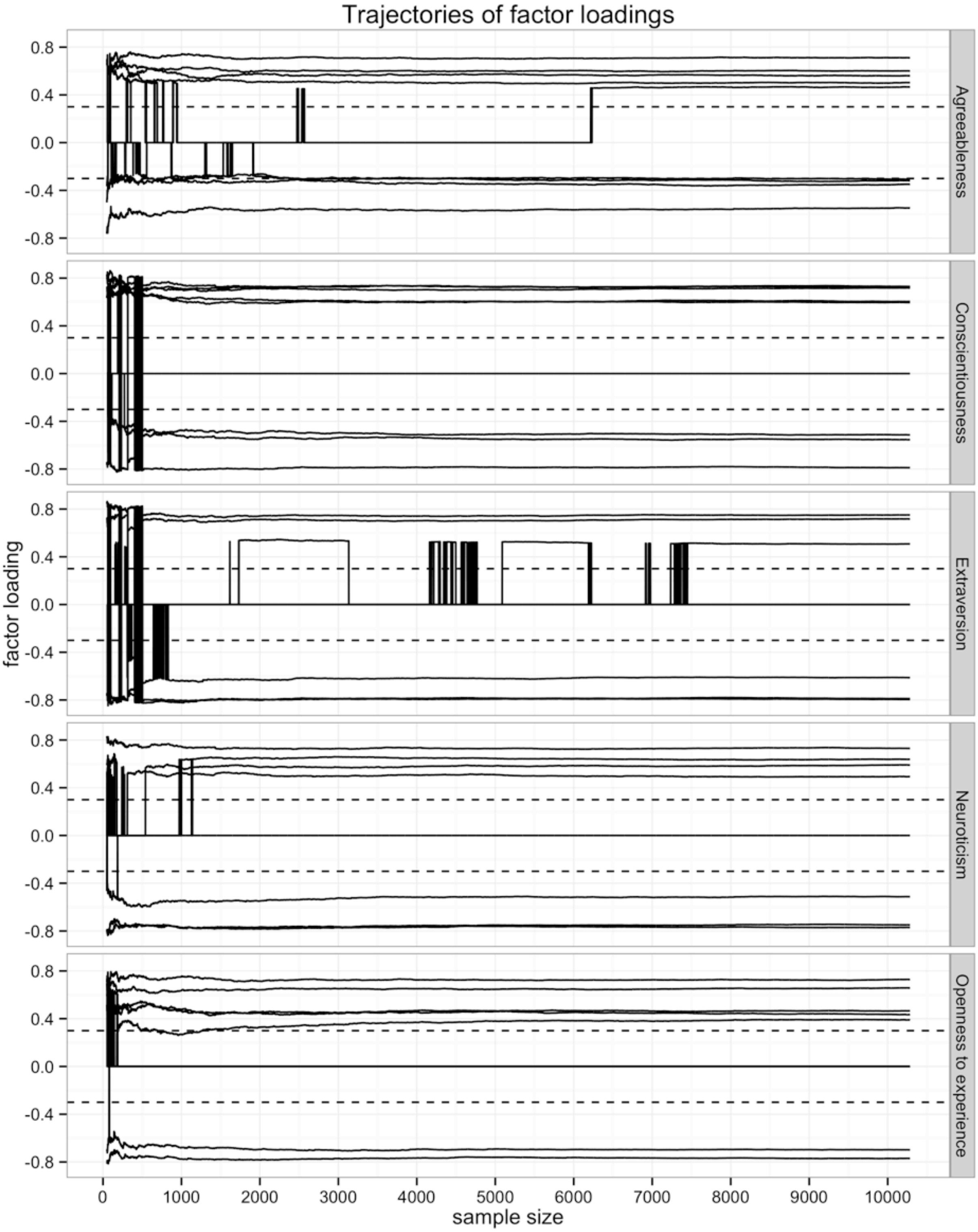

Figure 1. Primary factor loadings for individual items in sequential samples of participants.
Note: Horizontal lines represent the trajectories of primary factor loadings for individual items. Vertical breaks in these trajectories result from the simple structure criterion that was implemented. According to this criterion loadings were set to zero when primary loadings were smaller than .3 or cross-loadings larger than .3

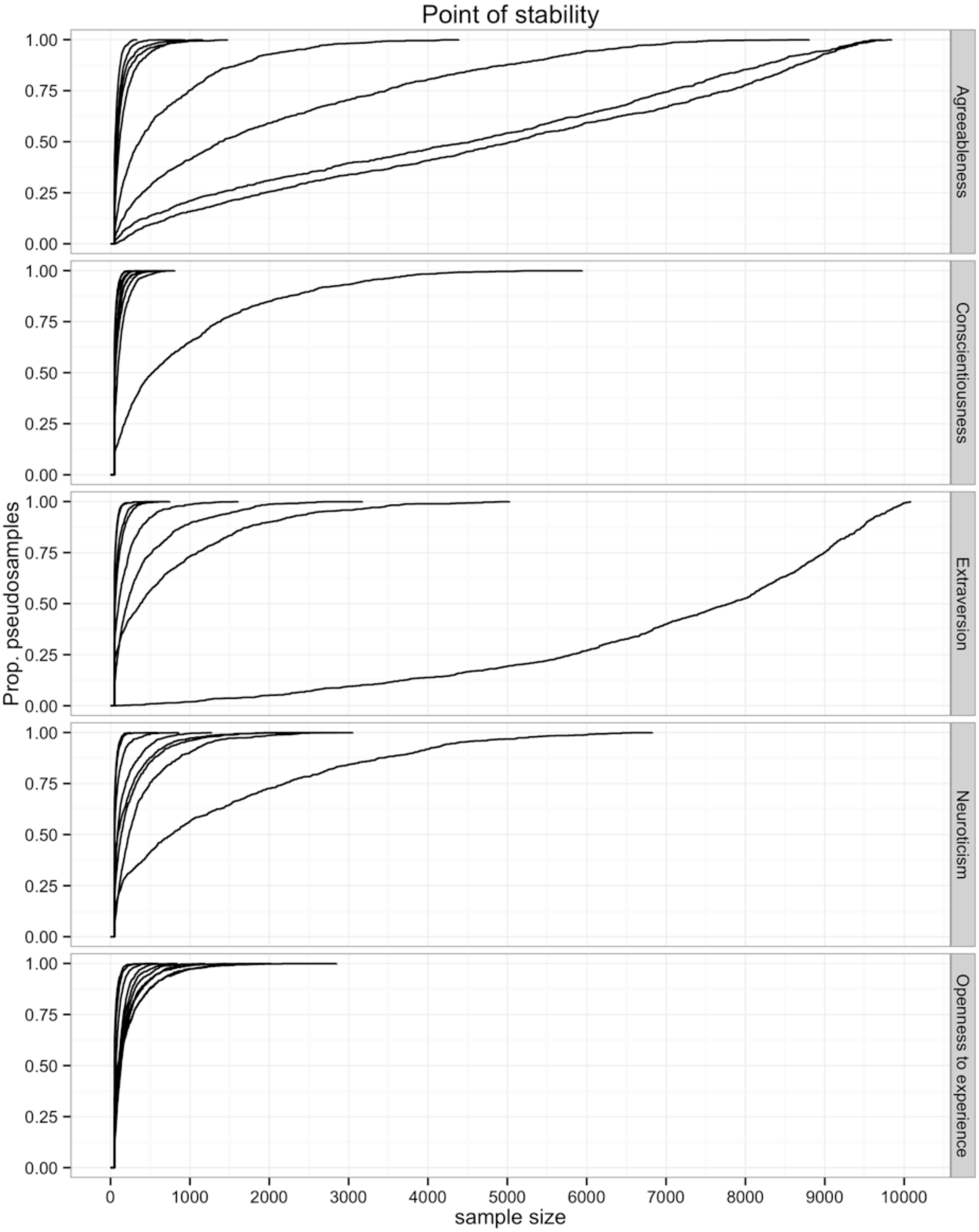

Figure 2. Cumulative frequency plot of the point of stability in different random orders.
Note: Each line represents an individual item. For each sample size the proportion of random orders with a POS smaller than the sample size, i.e. proportion of orders that already converged, is given.

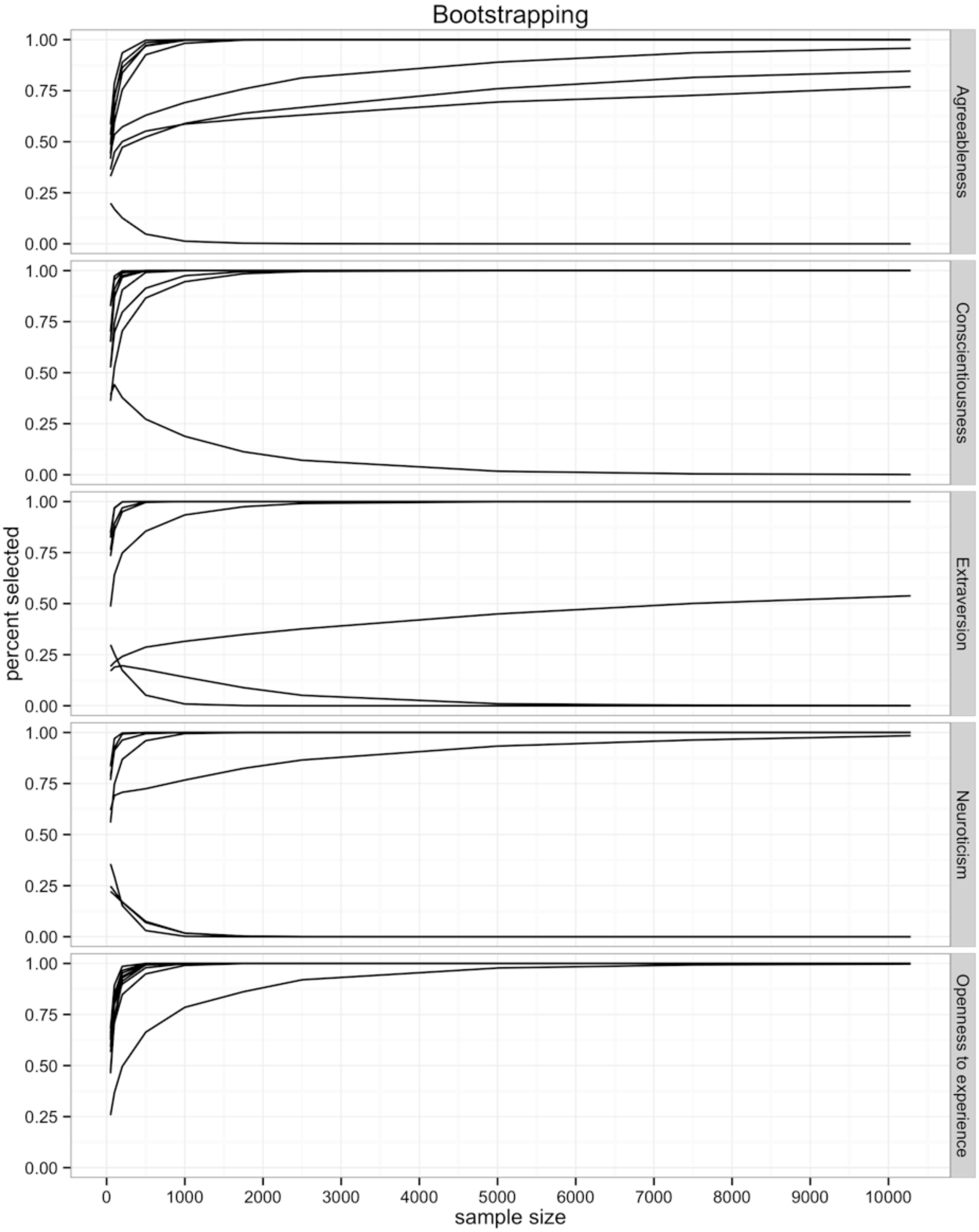

Figure 3. Results of the bootstrapping analysis.
Note: Each line represents an individual item. For each sample size the proportion of pseudosamples in which the item was retained is given.

2.3 Discussion

The results of the sequential sampling analysis indicate that the primary factor loadings stabilize only at large sample sizes (n > 500). Furthermore bootstrapping revealed that three items had inconsistent primary loading even in the largest sample size available. Items showing most variability, both in terms of sequential sampling and bootstrapping, were from the agreeableness and extraversion factors. Whether this is a general trend related to the construct these factors assess, or related to specific content or wording needs to be investigated in future studies (cf. Egan et al., 2000).

3 Study 2 – International Personality Item Pool Big Five measure

The second study investigates the stability of factor loadings in a publicly available dataset (http://www.personality-testing.info/_rawdata; accessed on 18.06.2012) comprising responses to the International Personality Item Pool Big Five measure (IPIP; Goldberg et al., 2006).

3.1 Methods

3.1.1 Subjects

Data were collected via an open survey, i.e. participants could visit the link (http://personality-testing.info/tests/BIG5.php) from anywhere and their primary motivation was to get their personality test results. Data collection began fall 2011 and ended June 2012. Participants were asked at the end of the questionnaire whether or not their data may be used for educational or research applications and only data of participants who agreed were used. Missing data was avoided during response and participants younger than 12 and older than 99 were removed, resulting in 11,065 complete responses. Remaining participants were between 13 and 87 years old (29 ± 12). About 49% (n = 5,455) of the participants were female.

3.1.2 Materials & procedure

Participants responded to a big five measure based on the International Personality Item Pool (Goldberg et al., 2006). The five factors (“Extraversion", "Agreeableness", "Conscientiousness", "Emotional Stability", and "Intellect/Imagination") are each measured by ten items. Participants responded using a 5-point likert scale ranging from 1=disagree to 5=agree.

3.1.3 Data analysis

The data was analyzed in the same fashion as before. All in all 11,065,000 (11,015 subsamples for each trajectory * 1,000 random orders + 50,000 bootstrapping samples) EFAs were computed in the second study

3.2 Results

3.2.1 Factor solution in the whole sample

The table of factor loadings for the full dataset (tab. 2) shows that all but four items had loadings larger than .3 on the proposed factor and no cross-loadings larger than .3.

Table 2. Factor loadings for full sample and points of stability replication sample.

	Factor loadings					Point of stability		
						10%	median[a]	90%
Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	percentile		percentile
q0	.714	-.056	.104	.008	.034	50	50	92.1
q5	-.696	.028	-.149	.042	-.019	50	50	118.2
q10	.677	-.259	.259	.131	.029	84	503.5	1557.7
q15	-.727	.143	-.056	.002	-.001	50	50	97
q20	.729	-.069	.255	.105	.06	50	165.5	852
q25	-.533	.103	-.166	0	-.261	62	389	1785.4
q30	.732	-.093	.18	.042	.03	50	50	166.1
q35	-.611	.051	.033	.066	-.05	50	50	91.1
q40	.646	-.06	.018	-.007	.129	50	50	122
q45	-.681	.156	-.083	-.03	-.016	50	50	108.1
q1	-.037	.048	-.506	-.009	-.073	50	73	492.9
q6	.385	-.045	.59	.045	.125	50	63	314.2
q11	.081	.248	-.394	-.153	.08	65	258	981.5
q16	.067	.06	.751	.058	.035	50	50	492.9
q21	-.156	.053	-.647	.012	-.027	50	77	492.9
q26	.023	.14	.587	.022	-.032	50	67	492.9
q31	-.309	.121	-.603	-.005	-.028	14.8	3260	782.2
q36	.157	-.063	.566	.11	.043	50	80	4552.9
q41	.132	.112	.694	.074	.076	50	74	492.9
q46	.378	-.159	.423	.128	.046	50	80	419.2
q2	.037	-.094	.029	.67	.041	50	52	492.9
q7	.041	.116	.034	-.542	.151	50	82	492.9
q12	-.018	.036	.081	.46	.225	50	135	617
q17	-.05	.356	-.063	-.548	.025	50	162	793.3
q22	.099	-.081	.079	.635	-.099	50	61	492.9
q27	.012	.178	.007	-.618	.084	50	77.5	492.9
q32	-.058	.082	.014	.584	.017	50	56	492.9
q37	-.032	.226	-.147	-.462	-.035	50	141	581.2
q42	.089	-.017	.078	.64	-.08	50	56	492.9
q47	-.01	.011	.041	.48	.243	50	248.5	4368.9
q3	-.117	.71	.07	-.057	-.073	50	50	94.1
q8	.106	-.559	.053	.022	.053	50	53	133
q13	-.144	.622	.145	.052	-.001	50	59	159
q18	.17	-.441	.006	.112	-.016	50	87	262.2
q23	-.08	.568	-.038	-.095	-.081	50	50	104
q28	-.061	.769	.02	-.055	-.088	50	50	75.1
q33	.015	.738	-.045	-.118	-.014	50	50	99.1
q38	-.017	.771	-.055	-.117	-.03	50	50	87
q43	-.049	.705	-.165	-.043	-.024	50	50	118.2
q48	-.246	.652	-.05	-.151	.065	50	128	651
q4	.022	-.015	-.03	.019	.593	50	50	99
q9	-.007	.191	-.007	.027	-.554	50	71	239.1
q14	.087	.139	.084	-.079	.553	50	69	157
q19	.041	.088	-.031	.108	-.428	50	89	257.1
q24	.214	-.086	-.008	.142	.622	50	89.5	302
q29	-.109	.047	-.107	.023	-.512	50	73	176.1
q34	.09	-.15	.005	.16	.509	50	75	183
q39	-.003	.055	-.102	-.047	.584	50	53.5	118
q44	-.141	.164	.144	.038	.427	50	107.5	28.2
q49	.225	-.021	.071	.033	.675	50	89	399.5

Note: [a] = fractions indicate ties in data

3.2.2 Point of stability in 1,000 random samples

The POS for the different orders were a little lower than in the first sample. Of the 50 items only 2 items had a median POS larger than 500 (Fig. 4). However, inspection of the 90% percentiles showed that 22 items had a POS larger than 500 and 16 items had POS larger than 1,000.

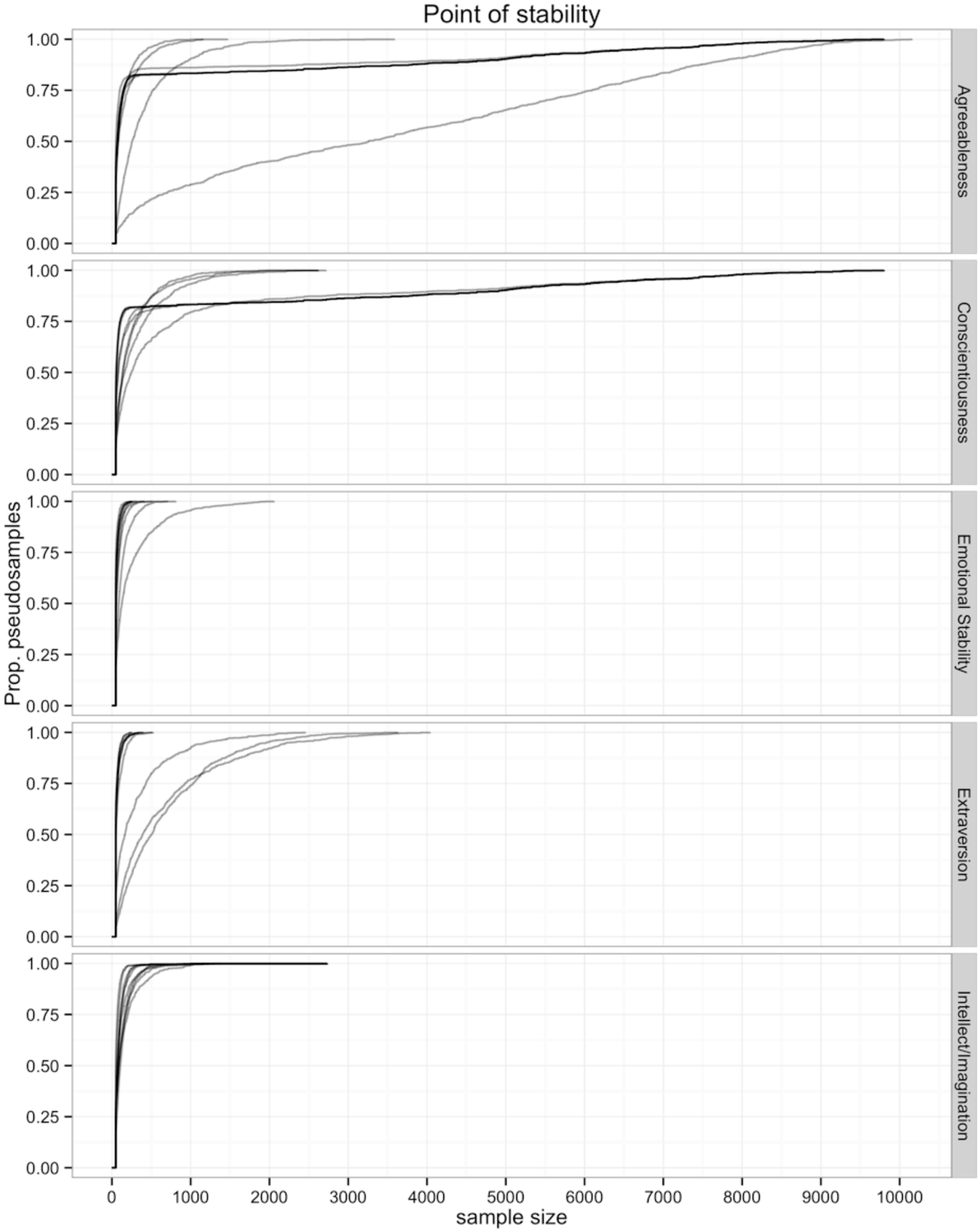

Figure 4. Cumulative frequency plot of the point of stability in different random orders study 2.
Note: Each line represents an individual item. For each sample size the proportion of random orders with a POS smaller than the sample size, i.e. proportion of orders that already converged, is given.

3.2.3 Bootstrapping results

The proportion of pseudo-samples in which an individual item was found to have a stable primary loading is depicted in figure 5. Seven items had unstable loading patterns with 500 participants and three items had unstable loading patterns with 1000 participants. Only one item from the agreeableness factor had an unstable loading pattern in the full data set.

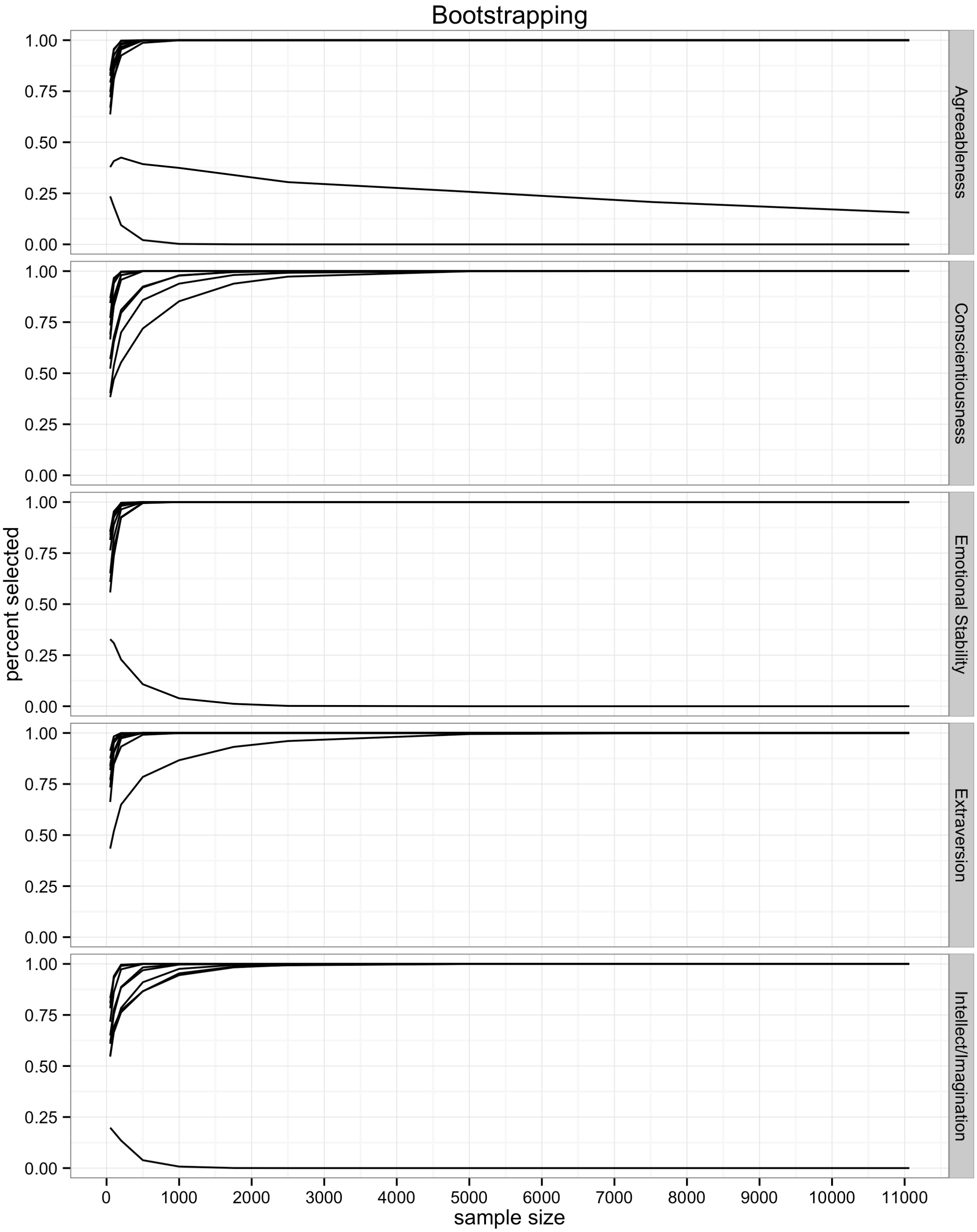

Figure 5. Results of the bootstrapping analysis study 2.
Note: Each line represents an individual item. For each sample size the proportion of pseudosamples in which the item was retained is given.

3.3 Discussion

Even though the primary factor loadings seem to stabilize at smaller sample sizes, the sequential sampling approach still identified many items that stabilize only at larger (n < 500) sample sizes. As in study 1 and earlier studies (Egan et al., 2000), the only item that was unstable in the full dataset was from the agreeableness factor. Since the IPIP uses a different item set and was presented in a different language than the BFI, specific wordings may be excluded as a reason for this pattern of results.

4 General Discussion

Factor loading patterns are widely used to develop and refine Big Five inventories. The aim of the present study was to establish the number of participants that are needed to yield stable loading patterns. In two datasets consisting of responses on widely-used measures of personality, we find that even sample sizes exceeding 500 or 1,000 participants which are large by conventional standards do not result in stable factor loadings for all items. Depending on the intended interpretation, it seems that at least about thousand participants are required to yield stable loading patterns for Big Five measures such as the BFI and IPIP.

Since our work is not based on simulated data we can only speculate about factors that give rise to the variability observed in the data. On the other hand, using actual rather than simulated data, our results give an estimate about how severe the consequences of variability of factor loadings is in applied research settings. Even though the amount of variability and the exact magnitude of the POS are different in the two datasets described here, both support the contention that item loadings are variable in typical datasets. Specifically, our findings suggest that inconsistent factor loadings for individual items between studies (Egan et al., 2000; McCrae & Costa Jr., 2004; Parker & Stumpf, 1998) may be due to chance fluctuations in the loading patterns.

In the present manuscript we have used two different methods to investigate the stability of factor loadings, sequential sampling, and bootstrapping. While both highlight the problem that factor loadings are sample specific, and that sample sizes below 500 participants are problematic to make decisions about the retention of individual items, they differ in the question they address and specific sample size they suggest. The sequential sampling approach addresses the question “would the results be the same if more data were collected?”. In a sense the criterion for stability according to the sequential sampling approach takes into account the results that would in larger sample sizes. In contrast to this bootstrapping addresses the question “how variable are the results at a specific sample size?”. As a result the sequential sampling approach may yield more conservative estimates of the stability of factor loadings and the sample size needed than the bootstrapping approach. Furthermore, the sequential sampling approach is computationally much more demanding than bootstrapping because the factor analysis is recalculated after each participant is added to the sample in order to generate the different trajectories from which the POS can be computed. Whether it is reasonable to perform the more complicated analysis depends on the research question. Researchers interested only in the variability of their results at a specific sample size, should consider bootstrapping to assess the stability of the factor loadings in their study.

A number of limitations need to be kept in mind when interpreting the results of the present study. First, both datasets were recruited online. While this may be more representative than the common college-student samples, it seems safe to assume that these participants may be more open to novel experiences than samples recruited by other means. Second, the results may be specific for several methodological characteristics, e.g. the cutpoint chosen for loadings, or the rotation method employed here. We have re-run our analysis using .4 as cutpoint for stability of item loadings and found similar results. As earlier simulation studies (e.g. Guadagnoli & Velicer, 1988) have shown similarities between different studies, we too believe that similar results would have emerged if we used a different rotational method. We would be most interested in any research adopting our method to study the variability of factor loadings in their sample. Third, as described above the sequential sampling approach implied that all items stabilize at some point. Accordingly this method is only applicable in very large sample sizes so that the distribution of POS can be scrutinized. In the datasets described here most loadings stabilize much earlier than the whole sample, giving some confidence that this was due to real stability rather than sample-size limitations.

Taken together, in two large datasets of widely-used measures of personality, we found that many primary item factor loadings do not stabilize until the sample size reaches about thousand participants. Further studies using the methods described here in other larger datasets and simulation studies are needed before decisive conclusions can be drawn about the precise sample size needed for factor loadings to stabilize. On a methodological level we believe that the utility of the sequential sampling approach is not restricted to correlation coefficients (Schönbrodt & Perugini, 2013), but is also useful for more complex multivariate procedures. Specifically, we believe that researchers should assess and report the variability of loading patterns since most studies will not have adequate sample size to yield stable loading patterns for Big Five measures such as the BFI and IPIP.

Acknowledgements

We would like to thank the partners of the PsyWeb project – Uwe P. Kanning, Fred Rist, and Stefan Schmukle – and Eric Jorgenson the maintainer of www.personality-testing.info for their support in using the data.

References

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research & Evaluation*, *18*.

Egan, V., Deary, I., & Austin, E. (2000). The NEO-FFI: Emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. *Personality and Individual Differences*, *29*(5), 907–920.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The big five inventory—versions 4a and 54. *Berkeley: University of California, Berkeley, Institute of Personality and Social Research*.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*, 102–138.

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen. *Diagnostica*, *47*(3), 111–121.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84.

McCrae, R. R., & Costa Jr., P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, *36*(3), 587–596. doi:10.1016/S0191-8869(03)00118-1

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Parker, W. D., & Stumpf, H. (1998). A validation of the five-factor model of personality in academically talented youth across observers and instruments. *Personality and Individual Differences*, *25*(6), 1005–1025. doi:10.1016/S0191-8869(98)00016-6

R Core Team. (2012). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*. Retrieved from http://www.R-project.org/

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0092656613000858

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). Using multivariate statistics. Retrieved from http://www.ulb.tu-darmstadt.de/tocs/135813948.pdf