Metodološki zvezki, Vol. 10, No. 1, 2013, 49-64

A First Look on Smaller Sized Samples for Bootstrap Derived Patterns of Profile Analysis via Multidimensional Scaling

Patrik P. Bratkovič[1]

Abstract

The possibility of using small sized samples was investigated for bootstrapping validation of scale values in Profile Analysis via Multidimensional Scaling (PAMS). Three original samples using three different psychological test batteries served as a basis for the investigation; TEMPS-A (N = 1167), BFQ (N = 347), and ICID (N = 565). Each of these samples were then randomly split into three smaller sizes (n = 50, n = 100, n = 200), and the original sample size (N = Full) was included as well. All four sample sizes were submitted to a bootstrapping procedure with 1000 resamples with replacement, and each bootstrapped resample was analyzed with Multidimensional scaling (MDS) to create two major profiles in PAMS. The resulting scale values, i.e. the coordinates from MDS, were analyzed using the bootstrapped distributions confidence intervals (CI). The smaller samples' CIs were compared towards the ones of the full sample to investigate invariance using Chebyshev's rule. The results indicate that the n = 200 samples were all invariant in comparison with the original sample sizes and produce reasonable results when the goal is to extract major profiles via bootstrapped confidence intervals using PAMS.

1 Introduction

Profile Analysis via Multidimensional Scaling (PAMS; Davison, 1996) is a technique based on non-metric multidimensional scaling (MDS). At its core it is a profile analysis which enables researchers to quickly get an overview of individuals' observed profiles, and compare them to what can be considered an underlying profile derived from the relations between all the profiles. The profiles derived by PAMS are most commonly validated through bootstrapped confidence intervals of the scale values (see e.g. Kim, Frisby, and Davison, 2004; Ding, Davison, and Peterson, 2005; Kim, 2010). This might however have an impact on how many participants a sample that will undergo validation of bootstrapped confidence intervals for PAMS scale values can have. What size the samples that would use a bootstrap procedure might have to be has not yet been investigated.

There are two reasons that the bootstrap is employed: firstly, the procedure of choosing scale values was seen as problematic since it was performed on the subjective basis that a chosen scale value was *probably* significantly different from zero (see e.g. Ding,

[1] Department of Psychology, University of Ljubljana, Ljubljana; patrik.bratkovic@gmail.com.

2001); secondly, there seems be an inherent lack of error theory involved with the application of MDS. Researchers in the field seem split between being content with the lack of statistical inference in applications of MDS, and others in the field would like to see this lack addressed (see Jacoby and Armstrong, 2012 for examples). This disparity has lead to attempts to bridge the gap, and a proposed non-metric maximum likelihood method of MDS does indeed have the possibility to produce standard errors without bootstrapping (see Takane and Carroll, 1981), unlike the method used in this paper. However, even though such methods might exist, there are very few substantial applied studies for them, and in the case of PAMS, a comparison with the bootstrap method indicated that the accuracy of the maximum likelihood method standard errors was significantly lower (see e.g. Kim, Frisby, and Davison, 2004; but also Weinberg, Carroll, and Cohen, 1984). The bootstrap use for PAMS confidence intervals has been the standard ever since and should a scale value's confidence band not include zero, it may be considered as salient and the combination of such salient scale values constitute a major profile (e.g. Kim, 2010). The bootstrapping procedure for confidence intervals in PAMS saw a major paper by Kim, Frisby, and Davison (2004) where they estimated the standard errors of scale values through bootstrapping – which in bootstrapping terms are the standard deviations. Recently, Kim (2010) investigated the invariance of PAMS profiles by splitting a very large sample and bootstrapping the scale values from the resulting 2 samples.

The aim of this paper is twofold; firstly, to investigate whether PAMS consistently extracts useful information across increasingly smaller sample sizes down to 50 – Chernick's suggestion for a smallest bootstrap sample size (2008). Hence, the aim is to evaluate whether extracted scale values for the smaller sample sizes can be considered salient, compared with the original, by submitting them to a bootstrapping procedure.

Secondly, we aim to evaluate the invariance of the profiles in the original sample and the smallest sample with matching salient scale values to find whether the profiles can be considered to be identified as a match.

Our overarching goal is hence to find a small sized sample for bootstrapped extraction of scale values where PAMS consistently provides reasonable and useful results.

2 Profile Analysis via Multidimensional scaling

The first, and for this paper most important, part of PAMS is the multidimensional scaling analysis. It is done on proximity data, distances between variable values, and there is no difference in the types of data that we could use compared to cluster- or factor analytic approaches. The analysis process itself however is different; the latter two analyses look at the individuals' proximity data and from there classify into clusters or factors that contain individuals, and from there estimate the underlying major profiles (Kim, Davison, and Frisby, 2007). PAMS on the other hand does it the other way around; it derives the profiles from the distances between subtests instead of individuals to estimate the relationships between derived and observed. Hence, it starts by estimating what the underlying patterns are, or what a "prototypical" person might look like based on the relations between the scores of the subtests across every individual, and then estimates how well the individuals scores on the subtests are in relation to the one of the prototypical person (Davison, Gasser, and Ding, 1996). The MDS approach thereby avoids the large sample

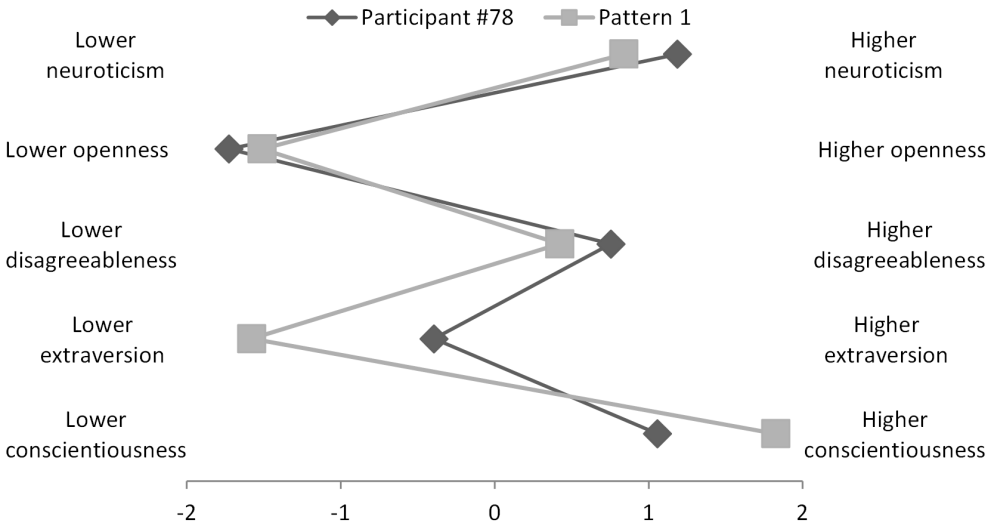

Figure 1: example of a PAMS major profile and the observed profile of a participant

problems and perhaps some small sample problems as well.

The PAMS model arranges differences between scores on scales into a pattern. The pattern is to represent the overarching characteristics of the test through the relations of the individuals' responses. Based on the participants' observed individual profiles, differences between the individual and the underlying pattern can be assessed. In Figure 1 above, the pattern indicates a contrast between conscientiousness and neuroticism on one side, and extraversion and openness on the other as a general profile pattern. It should immediately be noted that the profile lines connecting the scale values only serve to ease the visual comparison of the scale values between a major profile pattern and the participant's observed profile (see e.g. Ding, Davison, and Peterson, 2005), and do thus not indicate a functional shape - they can thereby be presented in the semantic differential chart format of Figure 1. In that respect it shows that participant #78 largely follows that major profile pattern, or rather that they do not differ much, as indicated by their observed scores. PAMS can thereby provide a quick overview of how a participant's observed profile matches the underlying characteristics of the test responses.

2.1 The PAMS model

The following equation defines the PAMS model:

$$m_{pt} = c_p + \sum_{k=1}^{K} \omega_{pk} \times \ x_{tk} + \varepsilon_{pt} \tag{2.1}$$

where:

m_{pt} is the observed score of an individual *p* on a subtest *t* where $t = 1, \ldots, T$. This is an element of a matrix where a row represents the individual *p* and the columns are for tests *t*.

c_p is a level parameter, which indicates the elevation of an individual *p*'s profile

$k = 1, \ldots, K$ is the amount of dimensions, in MDS terms, or profiles, in PAMS terms, that are extracted from the data matrix. Each profile *k* contains *T* scale values.

ω_{pk} is weight for how well the score of the profile for individual *p* aligns to a profile *k*.

x_{tk} is a test parameter, equaling the scale value *t* on a profile *k*.

Hence, the PAMS model divides the individual's observed test scores into two: firstly into a level, secondly into a pattern. The level is the expectation of the scores across all *T* tests, for the individual *p*, and thus determines the elevation of the observed profile for that individual. In the model, this is used to identify differences between individuals for their observed elevations. The scale values (x_{tk}) have to be noted as scores that handle the subtests and not the individuals; the weights (ω_{pk}) are instead associated with individuals in the sense that they note how well individual *p*'s observed scores match the prototypical pattern for *k*. For more detailed formulization of the PAMS model, see e.g. Kim, Frisby, and Davison (2004).

2.2 Identification conditions

To define the solution in the PAMS model, Davison (1996) added some assumptions and restrictions. Given $E(x_{tk}) = 0 \text{ for all } k$, it is implied that the major profiles, each of them, are ipsative – the sum of the scale values of each prototypical pattern amounts to zero. This bears the consequence of that the observed profile patterns could be reproduced, but that it goes beyond the prototypical pattern in that it has a level parameter (c_p). $E(\omega_{pk}^2) = 1$ for all *k* simply states the assumption that the squared individual weights are assumed to be 1. And equation $E(\omega_{pk} \times \omega_{pk'}) = 0$ for all $k, k' \neq k$ that the cross product sum between the person weights on two dimensions is equal to zero.

Put together, the above equations can be condensed to:

$$\delta_{tt'}^2 = \left(\frac{1}{P}\right) \sum_{p=1}^{P} (m_{pt} - m_{pt'})^2 \tag{2.2}$$

$$= \sum_{k=1}^{K} (x_{tk} - x_{t'k})^2 + 2\sigma^2$$

$$= d_{tt'}^2 + 2\sigma^2 \tag{2.3}$$

In Equation 2.3 it can be seen that $\delta_{tt'}^2$, apart from a constant $2\sigma^2$ equals the squared Euclidean distances $d_{tt'}^2$ across pairs of tests (t, t'). Thus, equation 2.3, yielded through 2.2, goes in hand with the statement that we are dealing with proximities from raw data, and that they will satisfy the fundamental assumption of non-metric MDS (see e.g. Ramsay, 1977; Takane, Young, and de Leeuw, 1977). This assumption states that the proximity measures, from the input data (m_{pt}), are monotonically related to the distances derived from the scale values (x_{tk}). Hence, if the data used satisfy the above model (each) one dimension will correspond to one prototypical pattern through the coordinates of the scale values along that one dimension.

2.3 The procedure

Given that the PAMS technique is largely unknown, the steps that constitute a PAMS procedure will be explained in this section. It is based on Multidimensional scaling, and begins with a non-metric MDS analysis on distance data created from the original multivariate matrix – individuals $\times$ objects. The MDS analysis provides a graphical representation of how dis/similar variables are in relation to each other. It provides spatial configurations with dimensions that the researcher must find sense in. These dimensions are often expressed as contrasts – e.g. east and west on a map representing a geographical map.

For a fictitious example, the proximity data one would use could be squared Euclidean distances across all pairs of subtests, for example in a battery designed to examine cognitive abilities. A dataset is fed to a MDS algorithm, a nonmetric scaling procedure (see Davison, 1996), where, for PAMS, *alternating least squares scaling* (ALSCAL, see Takane, Young, and de Leeuw, 1977) seems common, presumably due to its' ease of use and availability in the statistics software SPSS. To compute the distances, an individual $\times$ subtests score matrix serves as a basis for the computation (Kim, Davison, and Frisby, 2007). The MDS analysis yields at least one dimension, but most likely 2 or more. The positions of the variables as related to each other can be seen as coordinates. With this form of MDS we are interested in each dimension which is projected on a plane and corresponds to a major profile, and the coordinates are the scale values for each extracted dimension.

3 Why investigating smaller samples for PAMS?

PAMS technically does not have a "smallest" sample size suggestion one could turn to. In theory, it can use any sample size due to its basis in MDS (Davison, 1996). Arguably, this would however have to change provided that recent research has been promoting uses of bootstrapping techniques to confirm the profiles that it extracts. It extracts them using the dimensional coordinates yielded by MDS and projects them on a plane in order to create what is called a major profile – a combination of the coordinates, which are scale values in tests, indicating the characteristics of those who have responded. The confirmation of which scale values to use used to end here by looking at their values and determining which ones are "large enough" (See Kim, Frisby, and Davison, 2004). Nowadays however, if one wants a validation of which scale values would constitute a profile, this "first look" would however only hold for an initial look at what profiles a sample *might* include. To confirm the profile/s, bootstrapping of the scale values to find their confidence intervals has seen successful utilization (Kim, Frisby, and Davison, 2004). Since bootstrapping small samples could include biases (see Chernick, 2008), it would require an investigation of whether there is some minimum sample size which allows for reasonable results through the bootstrapping approach to confirmation of profiles. This could be immensely helpful for PAMS in general, given that it might find more use by students and research where temporal and economic factors could play a part in methods used for investigations.

4 Method

4.1 Data sets used

Temperament Evaluation of Memphis, Pisa, Paris and San Diego-auto questionnaire. TEMPS-A measures the affective temperament using depressive (DE), cyclothymic (CY), hyperthymic (HY), irritable (IR), and anxious (AN) subscales (see e.g. Preti et al., 2010). A total sample of 1167 participants, of which 63 % were female, completed the questionnaire in Slovenia (Dolenc, 2009).

Inventory of Child Individual Differences (or ICID, Slovene: VIRO). The ICID battery is valid across cultures and age, and tends to recover a Big Five factor like structure (Halverson et al., 2003; Župančič and Kavčič, 2004). It can retain 4 or 5 subscales; 5 were retained for this example. They are: conscientiousness (CO), openness (OP), extraversion (EX), neuroticism (NE), and agreableness (AG). A total of 565 participants' data was used after accounting for missing data (from 571), of whom 70 % were female (Vidmar, Grgić and Sočan, 2006).

The *Big Five Questionnaire (or BFQ)* was conducted by largely the same participants as the ICID questionnaire, but with 347 participants it was around 200 participants smaller than ICID (Grgić, Sočan and Vidmar, 2006). It includes the 6 subscales emotional stability (ES), extraversion (EX), conscientiousness (CO), agreeableness (AG), and openness (OP), and social desirability (SD) – which should allow finding potentially interesting links between SD and the other subscales.

4.2 Bootstrapped Confidence Intervals

We have 3 original, N = *Full,* samples (N = 1167, N = 565, and N = 347) that were split into 3 samples each, creating 4 samples for each questionnaire (N = Full, n = 200, n = 100, n = 50). These three smaller samples became pseudo-original samples in the sense that they served as the samples that were, in addition to the original sample, bootstrapped in the following tests. The reason for not using subsampling, a bootstrapping technique where one takes b samples of size n from a sample N where $n \ll N$ (see Politis, Romano, and Wolf, 1999), is that it might not portray what happens with a PAMS analysis with a small amount of participants. That is, if one would only collect data from 50 participants the subsampling procedure in this case might not accurately portray potential biases stemming from bootstrapping on 50 participants since with the subsampling, one would still sample on the entire group, just use a very small n each time. The current study examines the invariance of extracted profiles across various sample sizes and determines their properties statistically through bootstrapped confidence intervals.

Firstly, the original samples were submitted to a randomization procedure where 200, 100, and 50 participants in the sample were chosen without replacement to create one fixed sample of n = 200, n = 100, n = 50 for each test battery. The software generated 1000 bootstrap samples of each of those samples. Each bootstrap sample, for every sample size, was submitted to a MDS analysis using ALSCAL in SPSS (IBM, 2010). Thus each scale value, across sizes and tests, would have 1000 replicates, creating a bootstrapped sampling distribution. This procedure yields a mean, and a standard deviation of the

bootstrapped sampling distribution, which can be used to compute percentile bootstrapped confidence intervals – choosing the 2.5 and 97.5 percentile in the lower and upper tail of the sampling distribution provides us with a 95 % bootstrapped confidence interval for a scale value. Should that interval contain zero it can be assumed that the given scale value is not contributing to the contrasting nature of the PAMS solutions.

4.3 Test statistics

To qualify for the test statistic below on the samples, the sample's profile itself had to be similar to the original profile: the percentile intervals must indicate the same salient scale values across the subtests. For the profiles to be considered to match after qualification, the confidence intervals were tested for invariance. Using the null hypothesis *that there is no difference between the identification of scale values in two samples where one is the original (N = Full) sample*, the test procedure is as follows (using a random sample, here *n* = 200 as an example):

- First, widths of the confidence bands (*W*) for the original sample (*N = Full*) and the sample it is tested against *n* = (200; 100; 50) are computed; $W_{tk(A)} = x^U_{tk(Full)} - x^L_{tk(Full)}$, and $Wx^U_{tk(200)} - x^L_{tk(200)}$, with *U* meaning upper (97.5 %) and *L* meaning the lower (2.5 %) percentile value.

- Averaging the mean scale values across between the widths across all the scale coordinates, average of $W_{dif} = \frac{1}{T}\sum_{t=1}^{T}\left[W_{tk(N=Full)} - W_{tk(n=200)}\right]$

- Then the *Pooled Mean Standard Error* (PMSE) is estimated across the scale values where $PMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{bse^2_{tk(Full)} - bse^2_{tk(200)}}{2}\right)}$, the $bse^2_{tk(x)}$ referring to the squared bootstrap standard error of a scale value *t* on a dimension *k*, and *T* refers to the total amount of scale values *t* in a test.

- Lastly, the test statistic $c = \frac{Average\ of\ W_{dif}}{PMSE}$ is evaluated (see Kim, 2010). This is done since it cannot be guaranteed that the distribution of the test statistic would be standard normal, and the *c* is based on Chebyshev's rule which is thereby used for invariance comparisons. This variance measure states that a minimum amount of data can be found within *k* ($k > 1$) standard deviations of the mean, and it does not presuppose a normal distribution. Should the value of *c* be equal to or exceed |4.472|, more than 4.472 standard deviations away from the mean of zero, then the difference between confidence intervals will be considered significant and the profiles can no longer be considered to match. This is a quite conservative test, but given that normality cannot be assumed, it would be safe to use it for testing a statistical significant difference in our case (see also Kim, 2010).

5 Results

A two-dimensional solution was chosen for the three original samples. For ICID and TEMPS-A, the stress is ≈ 0.0000 and the RSQ is 1; BFQ's stress value is 0.0077 with

an RSQ of 0.99983. The following paragraphs show which sample sizes, if any, smaller than the original qualified for tests of invariance. The necessary qualification lies in that the profile patterns must match that of the original. By match it is assumed that to create the same profile as observed in the original; the smaller sample's patterns will include the same scale values as the original. The salience of the scale values is determined by the scale value's bootstrapped confidence bands not including the value 0.

TEMPS-A: in the first major profile, out of 5 sub scales the cyclothymic scale value had a bootstrapped CI which included 0, disqualifying it as salient and thus from being used in the creation of a profile (see e.g. Ding, Davison, and Peterson, 2005). There was a contrast between depressive, irritable, and anxious on one side and hyperthymic on the other. This pattern was repeated for the sample which contained 200 from the original sample, and the profiles could thus be considered similar. Smaller samples had more than the cyclothymic scale value's confidence bands that covered 0 and could not be considered similar; they did not qualify for invariance checks.

Major Profile 2 TEMPS-A: the second profile contains contrasts between depressive and anxious on one side, and cyclothymic and irritable on the other, with hyperthymic not being different from zero and thus not salient. Again, this pattern was obtained from the $n = 200$ sample, and the smaller samples did not succeed to replicate it (in fact, no scale value in the second major profile was salient for the $n = 100$ or $n = 50$).

Major Profile 1 BFQ: out of 6 subscales in the first profile of the BFQ we found that social desirability and to an extent emotional stability contrasts extraversion, conscientiousness, agreeableness, and openness. None of the scale values' confidence bands include 0; every scale value was included in the profile pattern for N = Full. This pattern is repeated for the sample with 200 participants, whereas the 100 and 50 samples' confidence bands are too wide to match those of the original sample; they include 0.

Major Profile 2 BFQ: Conscientiousness and openness are not salient in this profile. Extraversion and social desirability contrast emotional stability and agreeableness here, and do so for the sample with 200 participants as well. As indicated in previous examples, the 100 and 50 samples' confidence bands did not contain the same amount of salient scale values as N = Full and $n = 200$ (emotional stability and social desirability are the new non-salient scale values).

Major Profile 1 ICID: None of the scale values' confidence bands included zero and all were included in the profile for the original and the 200 sample. Other samples include at least one zero in their scale values' confidence bands and thus do not qualify for subsequent analysis by failing to match the original profile pattern. The profile is comprised of extraversion and openness which contrast conscientiousness, disagreeableness and neuroticism.

Major Profile 2 ICID: In the second major profile, conscientiousness and extraversion contrast neuroticism. Disagreeableness and openness are not salient in these profiles. These results are shared across the full sample and the sample with 200 participants. In the smaller samples, extraversion is not salient and they did thus not match the profile pattern of N = Full.

Complementary tests at $n = 150$ and $n = 175$ were made after suggestions from a reviewer and behaved similarly to $n = 100$ – they were inconsistent. While $n = 175$ showed relatively good characteristics it still showed inconsistencies in the production of scale values across all the test batteries; consistency was only found at $n = 200$ across

tests and it will remain as the sample size to be compared to the complete sample. This post hoc test is mentioned here to clarify the potential gap between n = 100 and n = 200 and will only be briefly mentioned again in the conclusions since it was not incorporated in the original hypothesis.

5.1 Evaluation of invariance across sample sizes in PAMS profiles

Tables 1 and 2 summarize the bootstrapping results for ICID, 3 and 4 summarize the BFQ results and the results for TEMPS-A are represented in tables 5 and 6. These tables include the N = Full and n = 200 samples given that the 200 participants sample was the only smaller sample which qualified for comparison with the full sample. The tables include: Dimensions 1 and 2 (means/coordinates from 1000 bootstrap replications in DIM1 and DIM2 respectively); Bootstrapped standard errors (BSE; standard deviation); the bootstrapped confidence intervals (BCI; lower (2.5 %) and upper (97.5 %) bands); and width whereas the average width difference, the pooled variance, and Chebyshev's rule (the test statistic denoted as "c" in the paper) are available as footnotes. This legend is repeated for table 1, including an explanation of the abbreviations for the scale values. The subsequent tables only include the abbreviations for the scale values. Results for each of the three (ICID, BFQ, and TEMPS-A) main samples and their 200 participants split samples follow in the upcoming paragraphs.

ICID: bootstrap confidence intervals were used to assess the invariance of the ICID profile patterns between the full sample and the 200 participants' sample. Given the null hypothesis that there were no differences between confidence bands between the original and the 200 sample, the Chebyshev's rule statistic c was computed for both major profiles. From table 1, we use the average width of the widths between the two samples, as the numerator of the test statistic c, and the pooled standard errors the denominator. This yields a ratio[2] of -1.684 indicating no significant difference between the bands.

Table 1: Descriptive statistics and Chebyshev's rule test statistic for the N = Full and n = 200 participants sample of ICID's first major profile.*

	N = Full				n = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
CO	1.826	0.063	1.684	1.948	1.821	0.110	1.597	2.022	-0.162	0.008
EX	-1.578	0.037	-1.654	-1.508	-1.576	0.059	-1.697	-1.466	-0.085	0.002
AG	0.423	0.033	0.358	0.495	0.423	0.055	0.313	0.530	-0.081	0.002
OP	-1.510	0.030	-1.569	-1.453	-1.507	0.051	-1.597	-1.405	-0.077	0.002
NE	0.840	0.072	0.702	0.996	0.839	0.126	0.598	1.080	-0.188	0.011

Source: Author's ad hoc processing on data of an ICID survey (Vidmar, Grgić and Sočan, 2006).
*Legend: Dimension 1; *CO*: conscientiousness; *EX*: extraversion; *AG*: agreeableness; *OP*: openness; *NE*: neuroticism; *BSE*: bootstrapped standard errors; *BCI*: bootstrapped confidence intervals; *lower*: lower 2.5 percentile; *upper*: upper 97.5 percentile; *W_dif*: the difference in width between the two samples BCIs; *var_p*: pooled variance.

[2] c = -1.684 through average W_dif = -0.119 and pooled standard error = 0.005 for the first dimension.

Similarly, testing for invariance in the second major profile was not significant either, indicating that a significant difference between the profiles scale values could not be found; they can be considered equivalent for now with a *c* of -2.017[3].

Table 2: Descriptive statistics and Chebyshev's rule test statistic for the *N* = Full and *n* = 200 participants sample of ICID's second major profile.*

	N = Full				*n* = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
CO	0.565	0.052	0.462	0.668	0.562	0.088	0.378	0.736	-0.152	0.005
EX	0.193	0.057	0.085	0.306	0.193	0.098	0.005	0.376	-0.150	0.006
DI	-0.004	0.045	-0.096	0.090	-0.007	0.078	-0.159	0.154	-0.127	0.004
OP	0.039	0.050	-0.054	0.139	0.039	0.087	-0.124	0.220	-0.151	0.005
NE	-0.793	0.040	-0.872	-0.715	-0.787	0.068	-0.924	-0.650	-0.117	0.003

Source: Author's ad hoc processing on data of an ICID survey (Vidmar, Grgić and Sočan, 2006).
*Legend: Dimension 2: *CO*: conscientiousness; *EX*: extraversion; *AG*: agreeableness; *OP*: openness; *NE*: neuroticism.

BFQ: the invariance of the BFQ profile patterns was assessed through bootstrapped confidence interval differences between the full sample and the 200 participants sample. Provided the same null hypothesis as for the previous test battery, stating no differences between confidence bands between the original and the 200 sample, *c* was computed for both profiles. From table 3, it can be gathered that with the average of the 6 scale value widths as the numerator, and with the denominator being the pooled standard error, we are presented with a ratio of -1.668[4] indicating no significant difference between the bands.

Table 3: Descriptive statistics and Chebyshev's rule test statistic for the *N* = Full and *n* = 200 participants sample of BFQ's first major profile.*

	N = Full				*n* = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
EX	0.393	0.041	0.310	0.475	0.399	0.054	0.288	0.509	-0.056	0.001
AG	0.815	0.054	0.711	0.916	0.815	0.071	0.677	0.948	-0.066	0.001
CO	1.067	0.055	0.953	1.175	1.057	0.068	0.926	1.188	-0.040	0.001
ES	-0.397	0.034	-0.461	-0.332	-0.395	0.045	-0.483	-0.301	-0.053	0.001
OP	1.000	0.041	0.921	1.080	1.001	0.053	0.891	1.100	-0.050	0.001
SD	-2.879	0.017	-2.910	-2.847	-2.878	0.020	-2.917	-2.837	-0.018	0.000

Source: Author's ad hoc processing on data of a BFQ test (Grgić, Sočan, and Vidmar, 2006).
*Legend: Dimension 1: *EX*: extraversion; *AG*: agreeableness; *CO*: conscientiousness; *ES*: emotional stability; *OP*: openness; *SD*: social desirability.

The *c* statistic for the second major profile indicated no significant difference between the differently sized samples' profiles – with a ratio of -1.191[5].

[3] with average W_dif = -0.140 and pooled standard error = 0.005 for the second dimension of ICID.
[4] with average W_dif = -0.047 and pooled standard error = 0.028 for the first dimension of BFQ.
[5] where the average of W_dif = -0.068 and the pooled standard error = 0.057 for the second dimension.

Table 4: Descriptive statistics and Chebyshev's rule test statistic for the N = Full and n = 200 participants sample of BFQ's second major profile.*

	N = Full				n = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
EX	-0.260	0.123	-0.528	-0.057	-0.280	0.144	-0.573	-0.037	-0.064	0.006
AG	0.521	0.079	0.357	0.644	0.507	0.102	0.309	0.650	-0.054	0.003
CO	-0.351	0.115	-0.491	0.024	-0.333	0.151	-0.520	0.132	-0.136	0.006
ES	0.255	0.059	0.124	0.375	0.265	0.085	0.087	0.408	-0.070	0.002
OP	-0.074	0.082	-0.235	0.090	-0.073	0.102	-0.260	0.129	-0.063	0.003
SD	-0.091	0.025	-0.135	-0.043	-0.087	0.031	-0.144	-0.031	-0.022	0.000

Source: Author's ad hoc processing on data of a BFQ test (Grgić, Sočan, and Vidmar, 2006).
*Legend: Dimension 2: *EX*: extraversion; *AG*: agreeableness; *CO*: conscientiousness; *ES*: emotional stability; *OP*: openness; *SD*: social desirability.

TEMPS-A: the profile patterns for TEMPS-A were assessed for invariance using bootstrapped confidence interval differences between the full sample and the 200 participants sample. Provided the same null hypothesis as for ICID and BFQ – there are no differences between the confidence bands of the original and the 200 sample – c was computed. Using the average of the 5 scale value widths as a numerator, and the pooled standard error as the denominator yields a ratio of -3.804[6] indicating no difference between the bands.

Table 5: Descriptive statistics and Chebyshev's rule test statistic for the N = Full and n = 200 participants sample of TEMPS-A's first major profile.*

	N = Full				n = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
DE	0.390	0.027	0.335	0.444	0.398	0.064	0.277	0.528	-0.142	0.001
CY	0.012	0.027	-0.041	0.066	0.013	0.070	-0.132	0.153	-0.178	0.001
HY	-2.598	0.013	-2.622	-2.570	-2.595	0.031	-2.657	-2.536	-0.069	0.000
IR	1.243	0.036	1.171	1.316	1.238	0.085	1.071	1.405	-0.188	0.002
AN	0.953	0.037	0.877	1.024	0.946	0.087	0.767	1.114	-0.200	0.002

Source: Author's ad hoc processing on data of a TEMPS-A test (Dolenc, 2009).
*Legend: Dimension 1: *DE*: depressive; *CY*: cyclothymic; *HY*: hyperthymic; *IR*: irritable; *AN*: anxious.

The c statistic for the second major profile was not significant either, indicating no significant difference between the profiles with a ratio of -4.016[7]. It should be noted that this value is relatively close to the limit of being significant using the c statistic - however, it should also be noted that it never passed this limit across many attempts and the cases here present the worst case scenarios we managed to find.

Using Chebyshev's rule, a rather conservative measure, N = Full and n = 200 were

[6] with the average of W_dif = -0.129 and pooled standard error = 0.034 of the first dimension.
[7] for the second dimension of TEMPS-A, the average W_dif = -.180 and the pooled standard error = 0.045, yield a ratio of -4.016.

Table 6: Descriptive statistics and Chebyshev's rule test statistic for the N = Full and n = 200 participants sample of TEMPS-A's second major profile.*

	N = Full				n = 200					
			BCI				BCI			
	mean	BSE	lower	upper	mean	BSE	lower	upper	W_Dif	var_p
DE	-0.286	0.045	-0.366	-0.205	-0.294	0.097	-0.466	-0.078	-0.228	0.002
CY	0.297	0.049	0.189	0.392	0.302	0.119	0.068	0.532	-0.260	0.003
HY	0.001	0.014	-0.035	0.021	-0.004	0.035	-0.097	0.047	-0.089	0.001
IR	0.475	0.042	0.387	0.563	0.454	0.146	0.217	0.657	-0.263	0.004
AN	-0.486	0.039	-0.561	-0.415	-0.457	0.134	-0.624	-0.239	-0.239	0.003

Source: Author's ad hoc processing on data of a TEMPS-A test (Dolenc, 2009).
*Legend: Dimension 2: *DE*: depressive; *CY*: cyclothymic; *HY*: hyperthymic; *IR*: irritable; *AN*: anxious.

consistently considered equivalent in the sense that they were extracting the same patterns through bootstrapped confidence intervals.

6 Summary and conclusions

PAMS can be useful in uncovering major profile patterns from a population. Observed participants' patterns can then be compared to these major patterns. One way of validating the patterns is through CFA (see Kim, Davison, and Frisby, 2007), or through the use of bootstrapped confidence intervals (see e.g. Kim, 2010). For invariance, Kim (2010) used the bootstrap approach to evaluate the matching of the profile patterns and their confidence bands in samples of similar, very large, sizes. For this paper, we used three different sample sizes (N = 347, N = 565, N = 1167) to evaluate the invariance of smaller samples (n = 50, n = 100, n = 200) and the original sized sample of the psychological test batteries, ICID, BFQ, and TEMPS-A, in an investigation of whether a smallest useful sample size could be uncovered.

6.1 A smallest sample size

Three samples that could potentially pass were used for the assessment in this paper (n = 50, n = 100, and n = 200). We proposed the use of invariance to assess the structure of these samples as compared to the original sample size. To qualify for the assessment their salient scale values had to be the same as those of the original sample. The assessment was based on the differences in range of the bootstrapped confidence bands from 1000 resamples, and tested via Chebyshev's rule. The results indicate invariance only between the original (N = Full) sample, and the sample with 200 participants across all three test batteries. This indicates that n = 200 should provide reasonable results and can be recommended, at the moment, for PAMS solutions using psychological test battery data when bootstrapping is used to determine the scale values, and thus the profiles.

Given that ICID, BFQ, and TEMPS-A are well-tested psychological batteries, the results were consistently similar across the samples. While the result for TEMPS-A was

close to the limit of the c statistic, it never passed it across several attempts, and it should again be noted that Chebyshev's rule provides a very conservative test. Yet, it could definitely warrant a future investigation. However this might not be the biggest issue for choosing the smallest sample size, instead it could be that by using percentile confidence intervals, the usual standard for PAMS, none of the smaller sample sizes even managed to qualify for invariance tests. The qualifications could thus be potentially problematic for the approach itself, and some way to correct potential bias in the distributions of small sized samples could be investigated in the future. A possible strategy to amend one culprit when it comes to bias, the CIs, would be to investigate whether there are CIs that would address this potentially problematic side of PAMS solutions as this could also potentially affect the width difference between the sample sizes. There could also be other studies done to investigate the type of resampling schemes that are used in PAMS and their impact on the solutions. That said, the $n = 200$ TEMPS-A did indeed indicate potential problems with the ratio being relatively near the limit of what could be considered significantly different using Chebyshev's rule, however it should be remembered that the reported value was the worst found across several tests, and that it thus never passed the limit of the ratio. Thereby, with this conservative test in mind, a recommendation of using 200 participants would be quite safe to make for practitioners that want to use bootstrapping in determining which scale values to include in a PAMS profile. It could also save resources, especially for researchers and students who might not have the possibilities to collect data from a very large amount of participants - or have database access to previously used data.

Another limitation lies in the use of only three datasets, and three random samples from within them – two additional qualification samples of $n = 150$ and $n = 175$ were investigated post hoc after reviewer suggestions, however they did not pass qualification. The investigation could also potentially have started out with simulated data first, and have been followed up by real data to see whether its results would match the simulations. The instruments are however well researched and used, and should represent characteristics that are commonly found with the use of psychological batteries. Additionally, from a practitioners perspective it might be more comforting to immediately see results from real data for lesser known analysis techniques. On the other hand, there could however be characteristics in respondent data that were not accurately captured due to the psychological nature of these tests and resulting data. Arguably, those characteristics can be difficult to fully capture with simulations as well due to the structure that can be found in psychological test battery data. Consequently, while the results for a smallest sample to successfully use for finding scale values were very consistent throughout these three batteries, and probably would not change across other psychological instruments, they could be expanded upon for a future study which should include simulations of various characteristics that could go beyond the scope of those commonly found in psychological test batteries. Hence, any suggestions made in this paper are first and foremost for researchers using psychological test batteries, with the potential to be extended to other fields as well.

Arguably, another limitation is that the research hypothesis proposed here might seem unreasonably harsh given that the confidence intervals would likely change with N. Yet, it did not indicate to be a prominent problem here, given the qualification of which samples to retain. Without that qualification however, it is not unreasonable to believe that smaller samples would be accepted and feasible to use, however one would have to resort to

another type of qualification if one wants to avoid the subjective picking of "large enough" scale values as a qualification for testing. Following that line of thought, it might be wise to also switch to a less conservative test statistic, albeit the use of Chebyshev's rule has previously seen use in testing for invariance of PAMS solutions (see Kim, 2010), and should provide safety to the recommendation made in this paper. Nevertheless, this could very well present a possibility for further testing on smaller sized samples with different, and similar, qualifications of which scale values to include in a PAMS profile.

Provided that profile examining is gaining more and more interest, it is our hope that the indications of a sample size which produces reasonable results for PAMS can be of help for researchers and students worldwide and that they should feel relatively safe using a $n = 200$ sample for PAMS studies involving psychological test batteries. Hopefully this insight will be useful either for research on participants, or in deeper investigations on the smallest sizes where PAMS will produce reasonable results when bootstrapping is used to determine the profiles.

Additional files

The code used to perform the analyses presented in this paper is available as supplementary information on the journal's web page.

Acknowledgment

The author is grateful to Dr Gregor Sočan for comments and help. The author would like to extend a special thanks to the anonymous reviewers and the Editor whose insightful comments and suggestions helped improving and clarifying this paper.

References

[1] Chernick, M.R. (2008): *Bootstrap Methods: A Guide for Practitioners and Researchers.* 2^{nd}ed. Hoboken, New Jersey: J Wiley.

[2] Davison, M.L. (1996): Multidimensional scaling interest and aptitude profiles: Idiographic dimensions, nomothetic factors. Presidential address to Division 5, American Psychological Association, Toronto.

[3] Davison, M.L., Gasser M., and Ding, C.S. (1996): Identifying major profile patterns in a population: An exploratory study of WAIS and GATB patterns. *Psychological Assessment,* **8**, 26-31.

[4] Ding, C.S., Davison, M.L., and Peterson, A.C. (2005): Multidimensional scaling analysis of growth and change. *Journal of Educational Measurement,* **42**, 171-192.

[5] Dolenc, B. (2009): *Psihometrična analiza dolge verzije vprašalknika temperamenta TEMPS-A:* diploma essay. Ljubljana: Faculty of Philosophy, University of Ljubljana.

[6] Grgić, K., Sočan, G., and Vidmar, M. (2006): Hierarchical structure of personality. Unpublished raw data.

[7] Halverson, C.F. (2003): Culture, age, and personality development. Presented at the 11^{th} European Conference on Developmental Psychology, Milan, Italy.

[8] IBM Corp. Released (2010): IBM SPSS Statistics for Windows, Version 20.0, Armonk, New York: IBM Corp.

[9] Jacoby, W.G. and Armstrong II, D.A. (2013). Bootstrap Confidence Regions for Multidimensional Scaling Solutions" *American Journal of Political Science,* Forthcoming.

[10] Kim, S.-K. (2010): Evaluating the invariance of cognitive profile patterns derived from Profile Analysis via Multidimensional Scaling (PAMS): A bootstrapping approach. *Journal of International Testing,* **10**, 33-46.

[11] Kim, S.-K., Frisby, C.L., and Davison, M.L. (2004): Estimating cognitive profiles using profile analysis via multidimensional scaling (PAMS). *Multivariate Behavioral Research*, **39**, 595-624.

[12] Kim, S.-K., Davison, M.L., and Frisby, C.L. (2007): Confirmatory factor analysis and Profile Analysis via Multidimensional Scaling. *Multivariate Behavioral Research,* **42**, 1-32.

[13] Politis, D.N., Romano, J.P., and Wolf, M. (1999): *Subsampling.* New York: Springer Verlag.

[14] Preti, A., Vellante, M., Zucca, G., Tondo, L., Akiskal, K. and Akiskal, H. (2010): The Italian version of the validated TEMPS-A: the temperament evaluation of Memphis, Pisa, Paris, and San Diego. *Journal of Affective Disorders,* **120**, 207-212.

[15] Ramsey, J.O. (1977): Maximum likelihood estimation in multidimensional scaling. *Psychometrika,* **42**, 241-266.

[16] Takane, Y. and J.D. Carroll. (1981) Nonmetric Maximum Likelihood Multidimensional Scaling from Directional Rankings of Similarities." *Psychometrika*, **46**, 389-405.

[17] Takane, Y., Young, F.W., and de Leeuw, J. (1977): Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 267-276.

[18] Vidmar, M., Grgić, K., and Sočan, G. (2006): Confirmatory and exploratory factor analysis of the ICID on the student population. Presented at the 26^{th}International congress of applied psychology, Athens, Greece.

[19] Weinberg, S.L., Carroll, J.D., and Cohen, H.S. (1984). Confidence Regions for INDSCAL Using the Jackknife and Bootstrap Techniques." *Psychometrika*, **49**, 475-491.

[20] Župančič, M., and Kavčič, T. (2004): Personality structure in Slovenian three-year-olds: The Inventory of Child Individual Differences. *Horizons of Psychology,* **13**, 9-28.