Multivariate Behavioral Research, 46:58–89, 2011

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.544227

Psychology Press
Taylor & Francis Group

Robust Mokken Scale Analysis by Means of the Forward Search Algorithm for Outlier Detection

Wobbe P. Zijlstra, L. Andries van der Ark, and Klaas Sijtsma
Tilburg University

Exploratory Mokken scale analysis (MSA) is a popular method for identifying scales from larger sets of items. As with any statistical method, in MSA the presence of outliers in the data may result in biased results and wrong conclusions. The forward search algorithm is a robust diagnostic method for outlier detection, which we adapt here to identify outliers in MSA. This adaptation involves choices with respect to the algorithm's objective function, selection of items from samples without outliers, and scalability criteria to be used in the forward search algorithm. The application of the adapted forward search algorithm for MSA is demonstrated using real data. Recommendations are given for its use in practical scale analysis.

Exploratory Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002) is a popular method for identifying scales from a larger set of items. Recent examples of MSA are found in criminology (e.g., Santtila et al., 2008), health sciences and medicine (e.g., Watson, Deary, & Shipley, 2008), marketing (e.g., Paas & Sijtsma, 2008), political science (e.g., Jacoby, 2008), psychiatry (e.g., Bech, Wilson, Wessel, Lunde, & Fava, 2009; Korner et al., 2007), psychology (e.g., Watson, Roberts, Gow, & Deary, 2008), and sociology (e.g., Loner, 2008). MSA analyzes discrete item scores that often equal 0 and 1 for incorrect and correct answers, respectively, or 0, 1, 2, 3, 4 for the degree of endorsement to the statements included in rating scale items. MSA may be hampered by the

Correspondence concerning this article should be address to Wobbe P. Zijlstra, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: w.p.zijlstra@uvt.nl

presence of outliers. The purpose of this study was to propose a method for dealing with outliers such that more trustworthy scales result from an MSA.

Outliers are observations or subsets of observations, which appear to be inconsistent with the remainder of the data (Barnett & Lewis, 1994, p. 7) and which may bias the results of a statistical analysis. For continuous variables, outliers usually are extremely large or small values, which are relatively easy to detect. Much research has been done into outlier detection for continuous variables and variables having many discrete categories (Atkinson & Riani, 2000; Barnett & Lewis, 1994; Chambers, Hentges, & Zhao, 2004; Rousseeuw & Leroy, 2003) but not into variables such as item scores having only a few different scores. Because the presence of outliers in data may result in biased statistical results and wrong conclusions (e.g., Barnett & Lewis, 1994, p. 317; Rousseeuw & Leroy, 2003, pp. 5–6), methods for dealing with outliers in item scores are badly needed. One possibility is to investigate the influence of outliers by analyzing the test or questionnaire data once including the outliers and once excluding the outliers. The result of this comparative analysis may be that outliers exercising a disproportionate influence on the statistical results are deleted from the data and that the remaining data are the basis for the definitive data analysis. Another possibility is to accommodate the outliers by constructing statistics that are robust with respect to outliers.

Defining outliers as a 0 score or as a 1 score among only two (or three, four, or five) possible values is not feasible. Alternatively, Zijlstra, Van der Ark, and Sijtsma (2007, 2011) argued that the definition of an outlier should be based on the vector of the scores on all the items in the test or the questionnaire that measure the same attribute. This way, in a group of respondents, most of whom produced varying item scores on a set of 5-point rating scale items, one respondent or a subgroup of respondents who produced, say, only 2s may be identified as suspect. This could not be accomplished by looking at individual item scores only because then a 2 score does not appear extreme. Defining the entire item-score vector as the observation of interest, Zijlstra et al. (2007, in press) found that the presence of outliers in test and questionnaire data may bias the scalability coefficients used in MSA. Hence, we expect that outliers also bias the selection of items produced by an MSA. As MSA ignores the possible presence of outliers in the data, this study aimed at identifying the outliers so as to produce an MSA result that is free of the influence of outliers.

In this study, we used the *forward search algorithm* (Atkinson, 1994; Atkinson & Riani, 2000; Atkinson, Riani, & Cerioli, 2004) to identify outliers in MSA. The forward search identifies observations that are inconsistent with the remainder of the data or have a disproportionate influence on interesting results such as parameter estimates, goodness-of-fit statistics, and significance tests. Identifying and removing the outliers facilitates assessing the magnitude of the bias. The forward search has been proven useful in regression models and generalized linear models (Atkinson & Riani, 2000); cluster analysis, discriminant analysis, and principal component analysis (Atkinson et al., 2004); factor analysis (Mavridis & Moustaki, 2008); analysis of variance (ANOVA) (Bertaccini & Varriale, 2007); and models for a mixture of categorical and continuous data (Cheng & Biswas, 2008).

In this article, we adapt the forward search to MSA. This adaptation involves choosing the algorithm's objective function, determining a unidimensional scale that serves as the criterion in the objective function, and selecting interesting statistics to be monitored as the forward search is run. The application of the adapted forward search for MSA was demonstrated using real data to which artificial outliers were added.

MOKKEN SCALE ANALYSIS

MSA is based on an item response model known as the monotone homogeneity model (MHM; Mokken, 1971, p. 118; Sijtsma, Meijer, & Van der Ark, 2011; Sijtsma & Molenaar, 2002, pp. 22–23). The MHM is defined by the assumptions of *unidimensionality*, which means that one latent variable denoted θ underlies the scores on the set of J items; *local independence*, which means that the J item scores are independent given θ; and *monotonicity*, which means the following. Let X_j be the integer item score on item j and let $E(X_j|\theta)$ denote the expected conditional item score, also known as the item response function (IRF); then, under the MHM, $E(X_j|\theta)$ is monotone nondecreasing in θ. The MHM is important because it is an ordinal measurement model. Let the total score be defined as $X_+ = \sum_{j=1}^{J} X_j$, then X_+ can be used to order respondents on θ (Sijtsma & Molenaar, 2002, p. 22).

Briefly, MSA selects subsets of items from a larger set such that each subset identifies a different attribute that is measured on an ordinal scale in agreement with the MHM. We first describe the item selection and then the estimation of IRFs, $E(X_j|\theta)$. Both ingredients are important for the adaptation of the forward search to MSA. Some additional notation is the following: The J items in a scale have scores in the same range with $x \in \{0, \ldots, m\}$. Equivalent scoring avoids the problems that different items are differentially weighted and have a different interpretation for the same score. For respondent v, the total score is denoted X_{+v} and an item score is denoted X_{vj}. We also need a total score without item j; this is restscore $R_{(j)}$, which is defined as $R_{(j)} = X_+ - X_j$.

Automated Item Selection

The automated item selection procedure (AISP; Sijtsma & Molenaar, 2002, Chapter 5) selects items from a larger set into a scale. The AISP uses the scalability coefficient H (Sijtsma & Molenaar, 2002, Chapter 4) as the selection criterion. First we define coefficient H, then a scale, and finally the AISP.

Let $Cov(\cdot)$ denote the covariance, and $Cov_{\max}(\cdot)$ the maximum covariance given the marginal distributions of the item scores. For two items, indexed j and k, the scalability coefficient is defined as

$$H_{jk} = \frac{Cov(X_j, X_k)}{Cov_{\max}(X_j, X_k)}.$$

For item j, the scalability with respect to the other items in the scale is defined as

$$H_j = \frac{Cov(X_j, R_{(j)})}{Cov_{\max}(X_j, R_{(j)})}.$$

The scalability of a set of J items is defined as

$$H = \frac{\sum_j Cov(X_j, R_{(j)})}{\sum_j Cov_{\max}(X_j, R_{(j)})}.$$

Items form a scale (Mokken, 1971, p. 184; Sijtsma & Molenaar, 2002, pp. 67–69) if, for interitem correlation ρ_{jk}, and for a positive lower bound value c of item scalability coefficient H_j, (a) $\rho_{jk} > 0$, for all item pairs (j, k), and (b) $H_j \geq c$, for all items j. The MHM implies the first condition (Sijtsma & Molenaar, 2002, p. 51). Positive correlations constitute a necessary condition for all J items measuring the same attribute represented by the latent variable θ. The second condition ascertains the selection of items that each contribute to the accurate ordering of respondents by means of their total scores on the J items. It can be shown that the second condition implies that the overall scalability coefficient H exceeds c (i.e., $H \geq c$; Sijtsma & Molenaar, 2002, p. 58).

The choice of c allows the researcher to define a lower bound for accuracy of person ordering. Experience has shown that $c = .3$ is a practically useful value (Mokken, 1971, pp. 184-185; Sijtsma & Molenaar, 2002, p. 60), which has been adopted in all MSA software and real-data analyses. Lower c values typically result in the selection of items, which have relatively flat IRFs and contribute little to accuracy, and higher c values easily result in the rejection of many items having relatively steep IRFs. This leaves only the very best items that, however, may be too few to cover the attribute well. Hemker, Sijtsma, and Molenaar (1995) acknowledged the element of arbitrariness using $c = .3$ and recommended trying different c values letting the scale selection depend on the presence of particular patterns of scale results across the different c values. For the forward search, this would complicate the adaptation to MSA considerably, and we preferred to stay close to what most practical researchers do, which is use $c = .3$. It may be noted, however, that different c values tend to produce different MSA results, and this may result in different forward search results.

Following common practice, the AISP consists of the following steps (Sijtsma & Molenaar, 2002, Chapter 5). First, the item pair having the greatest, positive H_{jk} value that is significantly greater than 0 and also exceeds $c = .3$ constitutes the starting pair (provided such a pair exists). From the remaining $J - 2$ items, a third item u is selected, which (a) correlates positively with items j and k (Condition 1), (b) has $H_u \geq c$ with the items j and k (Condition 2), and (c) of all candidate items for selection satisfying Conditions 1 and 2 produces the highest overall scalability H value together with the items j and k. When such an item can be selected, the AISP proceeds finding a fourth item u', a fifth item u'', and so on until a next item satisfying Conditions 1 and 2 is not available anymore. Then the items constituting the scale have been selected. When different latent variables underlie different subsets of items, the AISP may find a second scale from the remaining items, a third, and so on.

Estimating the Item Response Function

Following Junker and Sijtsma (2000; also see Sijtsma & Molenaar, 2002, pp. 40–41), to estimate the IRF, $E(X_j|\theta)$, we first replace latent variable θ by its ordinal estimator $R_{(j)}$ and then use the mean conditional item score, $\overline{X}_j|R_{(j)}$, to estimate $E(X_j|R_j)$. (Junker and Sijtsma, 2000, showed for 0, 1-scores that using X_+ instead of $R_{(j)}$ produces a biased IRF estimate, and that for rating-scale scores $R_{(j)}$ may also fail; however, in the absence of a useful alternative, we use $R_{(j)}$ throughout.) Conditional item means $\overline{X}_j|r$ are computed for $r = 0, \ldots, m(J-1)$. The IRF estimate is known as the item-restscore regression. Sparse or empty restscore groups are often observed in the tails of the restscore distribution, especially in small samples. Molenaar and Sijtsma (2000, pp. 67–74) recommend joining such adjacent restscore groups until the resulting combined group exceeds a minimum group size, denoted *minsize*. The default value is a compromise between too few groups (by which a violation of monotonicity could be masked) and too many groups (by which the fraction in a group would be very instable). Let N be the sample size. The recommended defaults are $N/10$ for $N > 500$, $N/5$ for $200 < N \leq 500$, and $N/3$ for $50 < N \leq 200$. These are the values commonly used in MSA; hence, we use them here as well. Mean item scores are estimated in the joined groups.

FORWARD SEARCH ADAPTED TO MSA

The terminology used in outlier analysis can be quite deceptive, which is the reason to start defining it precisely before we discuss the forward search. Bar-

TABLE 1
Types of Observations Identified by the Forward Search

		Suspected	
		Yes	*No*
Influential	Yes	Suspected and influential	Influential
	No	Suspected	Normal

nett and Lewis (1994, p. 7) defined outliers as observations that appear to be inconsistent with the remainder of the data. Such observations can be considered unusual, extreme, or surprising; in short, they are *suspect*. Observations that have a disproportionate effect on the results of the statistical analysis are *influential* (e.g., Beckman & Cook, 1983). Table 1 shows the four combinations of an observation being suspected or not suspected and influential or not influential. We distinguish observations that are neither suspected nor influential, that is, *normal*, suspected only, both suspected and influential, and influential only.

We define an observation as the vector of J item scores for respondent v. We adapt the forward search to MSA such that normal observations are consistent with the MHM and are kept in the MSA. Suspect observations are inconsistent with the model but they do not influence statistical analysis results such as scalability coefficients. However, due to their inconsistency with the MHM it is uncertain whether the item scores and statistics based on these item scores can be trusted in analyses beyond the MSA, and the researcher is advised to remove suspect observations from the data. Finally, the latter two categories of observations are both influential, biasing the outcomes of an MSA, and the researcher is advised to also remove these observations from the data. It may be interesting to distinguish the latter two categories, and we do this henceforth.

To identify the four types of observations mentioned in Table 1 in an MSA, we define a measure, which expresses the distance of an observation to the MHM. For respondent v, we determine the squared distance between each item score and the corresponding IRF, and we add the squared distances across the J items. This is the residual for respondent v relative to the J IRFs, comparable to the residual in a common regression analysis (Atkinson & Riani, 2000; for different statistical models, see Atkinson et al., 2004, and Mavridis & Moustaki, 2008). The sum of squared residuals is

$$\epsilon_v^2 = \sum_j \left[X_{vj} - E(X_{vj}|\theta_v)\right]^2. \quad (1)$$

This is the objective function. The forward search is a procedure that uses the residuals in such a way that all four types of observations can be distinguished.

The forward search (Atkinson & Riani, 2000; also see Hadi, 1992) adapted to MSA has three steps.

1. Selecting an initial subsample of size n_0 from the complete sample of size N. We estimate the measurement model—here, the J IRFs—in the initial sample and compute the residuals for all N respondents, not only the ones included in the initial subsample.

2. Progressing the forward search. In the next step, we select from the N residuals the $n_0 + 1$ lowest values and assign the corresponding respondents to the subsample. It may be noted that the subsample does not necessarily consist of the n_0 respondents in the initial sample plus 1; instead, what happens regularly is that upon selection on the basis of the smallest residuals q respondents leave the subsample and $q + 1$ different respondents enter. We reestimate the J IRFs in the subsample of size $n_0 + 1$ and compute the residuals based on these new estimates for the complete sample. The forward search progresses by repeating this step—select a new subsample that contains one observation more than the previous subsample, reestimate the J IRFs, and compute the new residuals for the complete sample—until all N respondents have been selected and the subsample equals the complete sample. Atkinson and Riani (2000) showed that in the earlier steps, due to their large residuals suspect observations are quickly pushed out of the subsample and nonsuspect observations are included, yielding a purified subsample with respect to the model of interest. Observations with large residuals are expected to join the subsample in the final steps.

3. Monitoring the forward search. As the forward search progresses, for each observation it is checked how the residual develops and also how relevant statistical outcomes change with increasing subsample size so that suspect and influential observations can be identified. In MSA, the partitioning of the item set into scales by the AISP, the estimated IRFs, the J item coefficients H_j, and the total-scalability coefficient H are excellent candidates to be monitored.

The forward search is run several times. The first forward search is used to find a partitioning of the items into unidimensional scales that is robust to influential and suspect observations. Once the scale structure of the data has been established, the forward search is run for each scale separately while monitoring its scalability coefficients and estimated IRFs.

Objective Function

Two problems are relevant when $\overline{X}_j | R_{(j)v}$ is used to estimate $E(X_{vj}|\theta_v)$ (Equation 1). The first problem is that parameter θ_v is estimated J times using varying restscores $R_{(j)v}$ across items; in particular, for items scored $0, \ldots, m$ the restscore can attain values $R_{(j)v} = X_{+v} - m, \ldots, X_{+v}$. For example, for $m = 4$ a respondent with $X_{+v} = 19$ can have restscores $R_{(j)v} = 15, 16, 17, 18, 19$. This variation causes respondent v to be a member of different (joined) restscore groups, and a pilot study showed that this variation caused the forward search to be unstable in the sense that respondents were found to be influential because they had different restscores for different items, not because they showed real misfit to the model. We wished to eliminate this methodological artifact.

Replacing the restscore by total score X_+ clearly remedies the methodological problem but introduces the problem of a biased estimated IRF, $\overline{X}_j|X_+$ (Junker & Sijtsma, 2000). It was suggested to us to circumvent this problem by either using the *mean* restscore, but this alternative suffers from the same problem as X_+ because $\sum_{j=1}^{J} R_{(j)} = (J-1)X_+$, or by using the *median* restscore, but here the problem may be that a median group represents a highly heterogeneous θ distribution. Moreover, different medians do not order latent variable θ. Thus, we resorted to X_+, and found that the use of $\overline{X}_j|X_+$ at least produced a stable forward search ruling out the emergence of artifactual influential observations that resulted when $R_{(j)}$ were used. Using X_+, sparse total-score groups were also joined until the resulting group contained at least *minsize* cases (cf. Molenaar & Sijtsma, 2000, pp. 67–70).

We distinguish using X_+ as conditioning variable for determining the residual in Equation 1 from using $R_{(j)}$ as conditioning variable for estimating the IRF in MSA, which produces the unbiased ordinal estimator $\overline{X}_j|R_{(j)}$. In determining residuals for the forward search, our goal was to prevent artifactual influentials from appearing in the analysis, and in MSA our goal was to have an unbiased ordinal estimate of the IRFs. Hence, the objective function used in the forward search was

$$e_v^2 = \sum_j \left[X_{vj} - \overline{X}_j|X_{+v}\right]^2. \qquad (2)$$

The second problem is that for the determination of the residuals (Equation 2) a unidimensional item set is required so that the IRFs relate to one latent variable θ, estimated in the forward search by total score X_+. In the first forward search this poses a problem because the dimensionality has yet to be determined. The researcher should thus choose a plausible item set for computing X_{+v} and e_v^2 in Equation 2. Usually, items have been constructed with the intention to measure a specific attribute, allowing the item set to be partitioned into hypothesized scales before the analysis. One may base the residuals in Equation 2 entirely on the items in the hypothesized scales.

Choosing the Initial Subsample

Initially, Atkinson and Riani (2000; also see Atkinson et al., 2004) advocated the initial subsample to be outlier free, but Atkinson and Riani (2007) found that the composition of the initial subsample did not influence the final results and noted that observations can be selected at random to have the same effect.

For MSA, we chose the initial subsample size n_0 on the basis of experience accumulated elsewhere, which suggested equating n_0 to the number of model parameters (Atkinson & Riani, 2000, p. 31; Atkinson et al., 2004, p. 65; Mavridis & Moustaki, 2008). Then, for MSA, let G_j be the number of restscore groups used to estimate the IRF of item j based on the entire sample. If none of the restscore groups have to be joined, then $G_j = m(J-1)$; otherwise $G_j < m(J-1)$. We chose $n_0 = J \times \min(G_1, \ldots, G_J)$. This size is large enough to obtain an estimate of the IRF, $\overline{X}_j | R_{(-j)}$, albeit an inaccurate one. Hence, n_0 observations were drawn at random to constitute the initial subsample.

Progressing the Forward Search

Except for the restriction that each restscore group G_j should contain at least one observation, we chose to put no additional restrictions on observations entering or leaving the forward search. Yet the following additional restrictions may be applied. Mavridis and Moustaki (2008) constrained the forward search such that the next subsample contained the n observations from the previous subsample plus one new observation. They expected this procedure to facilitate the monitoring step. However, a drawback is that influential observations included in the initial subsample cannot leave it and affect the statistical analysis in every next step. This may induce masking and swamping (e.g., Hadi, 1992). Atkinson et al. (2004, p. 306) proposed a constrained balanced forward search in which the distribution of an available grouping variable in the subsample is proportional to its distribution in the entire sample. This procedure has been used in discriminant analysis (Atkinson et al., 2004, Chapter 6) and ANOVA (Bertaccini & Varriale, 2007), but it is unclear whether it has advantages over the regular forward search.

Monitoring the Forward Search

As the forward search progresses, the *forward plot* is helpful to monitor the objective function and the statistical results of interest in an effort to identify suspect observations and influential observations. Atkinson and Riani (2000) monitored interesting statistics in the context of regression analysis, Atkinson et al. (2004) did this for statistical analyses assuming multivariate normal data, Bertaccini and Varriale (2007) for ANOVA, and Mavridis and Moustaki (2008) for factor analysis.

The horizontal axis of the forward plot keeps track of the steps and is divided in three intervals. The first interval runs from n_0 to n_1 and identifies the first part of the procedure in which the forward plot is unstable and cannot be trusted. Atkinson and Riani (2007) found that an effective strategy for determining n_1 is to run the forward search B times, each time starting with a freshly drawn initial subsample, and then finding the value of n_1 at which the forward plots start to be highly similar. For MSA, "highly similar" was defined closely following Atkinson & Riani (2007, Figure 6). Let n denote the step at sample size n. For each step, we counted the number of unique subsamples out of B forward searches and identified the first step for which fewer than five unique subsamples were found. The sample size reached at this step was defined to be n_1. We chose the relatively small value of $B = 100$ because running the forward search is time consuming.

The second interval runs from n_1 to n_2, which is the first subsample size where observations with large residuals start entering the subsample. Depending on the function plotted, sharp changes are typical of influential observations and high values of suspect observations. Subsample size n_2 was determined using a method to be explained later. The third interval runs from n_2 to N, when suspect observations having large residuals join the subsample. Next, we discuss four types of forward plots used in a forward search for MSA.

Forward plot of objective function. Figure 1 shows the forward plot for four functions e_v^2 (Equation 1) as the subsample grows from n_0 to N (data explained in detail in next section). The four functions represent the four cells in Table 1. The solid curve represents a normal observation, which has a small residual throughout the procedure, reflecting good model fit. The dashed curve represents a suspect observation, which has a large residual throughout without becoming influential; it simply does not fit the model. The dashed-dotted curve represents a suspect observation that was also influential when it was included in the subsample at the end of the procedure between n_2 and N. This is a maverick, as one observation cannot really be expected to be influential among a sample in excess of, say, $N = 300$, typical of MSA applications. The dotted curve shows a case that is less suspect than the previous two until halfway n_1 and n_2 after which the curve suddenly decreases. This suggests that a cluster of observations simultaneously entered the subsample and drew the model toward them thus producing small residuals, which grew again as more normal observations entered the subsample. Thus, for large N influentials are expected to be identified in groups.

Minexcl-plot and maxincl-plot. The *minexcl*-plot (Atkinson et al., 2004, pp. 68–69) is the forward plot of the minimum residual (i.e., the best case) for the observations not in the subsample. We used the *minexcl*-plot as follows for determining the value of n_2. For the observations in the subsample, Tukey's upper fence (aka the boxplot upper fence; Tukey, 1977, pp. 43–44) for the corresponding residuals was computed and plotted. The value of n_2 is the size of the smallest subsample for which the minimum residual for the observations not in the subsample (i.e., *minexcl*) exceeds Tukey's upper fence. The rationale for this procedure is that observations in the subsample are considered non-

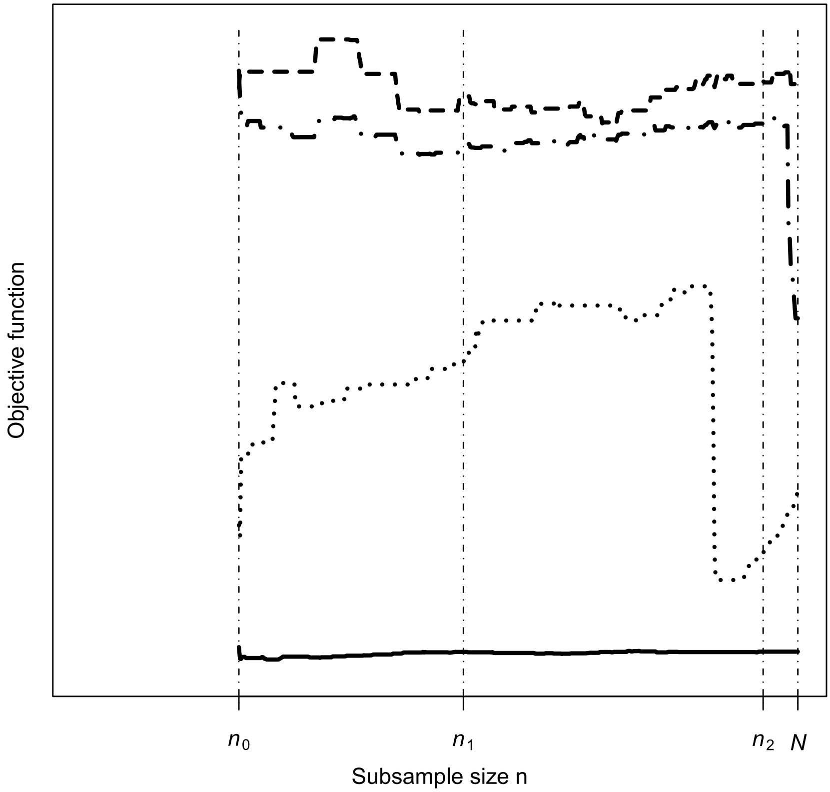


FIGURE 1 Forward plot of objective function (e_v^2) for normal observation (solid curve), suspect observation (dashed curve), suspect observation that is influential (dashed-dotted curve), and influential observation (dotted curve).

suspect, and Tukey's upper fence indicates the upper bound for residuals that may be encountered if the subsample is indeed free of suspect observations. However, if the *minexcl*-plot exceeds Tukey's upper fence, then even the best-fitting observation not in the subsample is considered suspect. This implies that all observations not in the subsample can be considered suspect. The *maxincl*-plot is the forward plot of the maximum residual (i.e., the worst case) for the observations in the subsample. It is defined similar to the *minexcl*-plot.

Both plots increase as the subsample grows. Suspect observations have high residuals and enter the subsample in the final steps, causing a sharp increase in the *minexcl*-plot and the *maxincl*-plot between n_2 and N. When an observation enters the subsample and strongly influences the objective function, it may cause observations not in the subsample to have lower residuals, which is visible from a downward peak in the *minexcl*-plot. Simultaneously, observations in the subsample may have higher residuals, which is visible from an upward peak in the *maxincl*-plot. The plots do not provide important information on normal observations.

Gap plot. The *gap plot* (Atkinson et al., 2004, p. 69) shows the difference between *minexcl* and *maxincl* (i.e., *minexcl* − *maxincl*). This difference is the minimum gap between residuals in the subsample and residuals in the remainder of the sample. The gap plot is a sensitive tool for detecting influential observations that are not suspect. An influential observation causes *minexcl* to decrease and *maxincl* to increase and, as a result, the gap becomes negative and the gap plot shows a downward and negative peak between n_1 and n_2. Influential observations enter the subsample just prior to the downward peak. The forward plots of the objective function for the individual observations entering and leaving the subsample (Atkinson et al., 2004, pp. 392–395) at these steps are used to identify the influential observations. For brevity, we refer to these plots as *follow-up plot-n*. Follow-up plots showing a sharp decrease (see Figure 1, dotted curve) identify influential observations.

Heuristics for identifying the sample prior to the downward peak in the gap plot are unavailable but we recommend starting at least two or three steps before the gap plot starts descending toward the peak. Influential observations detected using the gap plot should be removed immediately due to their detrimental effect on the remainder of the procedure, and the forward search should be repeated without the influential observations.

Forward plot of statistical results. The forward plots of interesting statistical results are used to identify observations that strongly influence these statistical results. Sharp changes between n_1 and N identify the influential observations, which are then removed from the analysis. In the first forward search, the most interesting quantities to monitor are related to the partitioning of the items produced by the AISP. In the subsequent forward searches interesting quantities to monitor are scalability coefficients H_j and H and the estimated IRFs ($\overline{X}_j | R_{(-j)}$).

A plot showing the number of scales for each subsample and item-entry plots are particulary useful to inspect the dimensionality of the data. Item-entry plots have the items listed on the vertical axis. For a particular scale, the item-entry plot shows whether or not the item belongs to the scale. As the forward search progresses, observations entering the subsample have larger residuals and are more deviant to a unidimensional model. As a result, the number of scales is expected to increase.

In the initial random subsample, the H_js and H are expected to deviate only by chance from the total sample results, but in the first few steps as large residuals are replaced by the smallest ones, the coefficients increase. As the subsample grows, observations entering the subsample have larger residuals, which has the effect of lowering the H_js and H.

For the forward plot of the estimated IRF, the composition of the restscore groups of an item must be constant as the procedure progresses. This was established by using restscore groups throughout that had been computed for the whole sample.

DATA EXAMPLE: AUTONOMY-CONNECTEDNESS SCALE

If one knew the truth, it would be possible to distinguish *regular* observations from *contaminant* observations; regulars come from the population of interest and contaminants come from another population. In real-data analysis, the truth is unknown and one cannot know for sure whether an observation is a regular or a contaminant. However, what one can do is distinguish normal observations from suspect observations and influential observations and hope that they match the regulars and the contaminants, respectively. We analyzed a real-data set in which by definition only the distinction between normal, suspect, and influential observations is possible. We added 31 artificial contaminants of which one was suspect and the other 30 influential. This enabled us to study the effectiveness of the forward search for identifying normal, suspect, and influential observations.

Method

Data. The subscale "sensitivity to others" from the short form of the Autonomy-Connectedness Scale (ACS; Bekker & Van Assen, 2006) contains 17 Likert items ($x \in \{0, \ldots, 4\}$) (Appendix A). The subscale measures the sensitivity to the opinions, wishes, and needs of other people; empathy; and the capacity and need for intimacy and separation. The sample without the artificial contaminants consisted of 588 students from Tilburg University who completed the inventory as a mandatory task in a psychology course. Due to the compulsory nature of the task, some students may have been obstinate, thus responding dishonestly or inaccurately, and their unusual item-score vectors may influence MSA results. There were no missing values. The mean total score was $\overline{X}_+ = 43.8$ ($SD = 8.8$).

Most constructs measured in behavioral research are defined by psychological theories, which do not pertain to properties relevant for measurement such as unidimensionality and monotonicity. Therefore, it cannot be expected that a set of items representing a construct in a psychological sense automatically satisfy the criteria required for measurement. This was also true for the subscale "sensitivity to others" that, according to Mokken's criteria, did not form a scale: $H = .25$. The aim of MSA was to construct a unidimensional scale consisting of as many items as possible so as to remain as close as possible to the psychological construct measured by the test.

To evaluate the usefulness of the forward search, the 31 artificial contaminant observations were added to the data ($N = 619$). The suspect observation (#619) stood out from the others; for this observation the scores on the negatively worded items were not recoded. The scores of 30 influential observations (#589, ..., #618) were affected by an extreme response style: they had high scores (3 or 4) on the eight most popular items and low scores (0 or 1) on the nine least popular items. All influential observations had the same, relatively low, total score ($X_+ = 37$) but their response patterns were unique. As a result, they also had low rest scores.

Analysis. The analyses consisted of a first forward search analysis to determine the dimensional structure of the data, in which we monitored the item partitioning, and a second forward search analysis on the largest unidimensional item set in which we monitored the IRFs and the scalability coefficients. Because the 17 items were intended to measure the same construct, we used all items to compute the residuals (Equation 2) for the first forward search. Both forward search analyses consisted of (a) determining n_0 and n_1, (b) identifying the influential observations, and (c) determining n_2 and identifying suspect observations. Influential observations found in Step b were removed and Steps a and b were repeated until no additional influential observations were left. After that, we moved to Step c, identifying suspect observations. The analyses were done using the R package `fwdmsa` (Zijlstra, 2010). The data can be obtained from this package. Appendix B shows the R code for the data analyses.

Results of the First Forward Search Analysis

Finding n_0 and n_1. The initial subsample size was $n_0 = J \times \min(G_1, \ldots, G_J) = 17 \times 8 = 136$. We found $n_1 = 292$ (Figure 2). Running 100 forward searches took approximately 40 min. All 100 forward searches yielded similar results between n_1 and N.

Inspecting forward plot of objective function. Figure 3 shows the forward plot of the residual, e_v^2. The 619 curves representing the observations cannot easily be disentangled but the software allows plotting or highlighting the objective functions of one or more observations of choice. The isolated, suspect contaminant observation (#619) shows up as the black dashed curve, which is largest for all steps (cf. Figure 1, dashed curve). The cluster of 30 contaminants is identified by the solid black curves (cf. Figure 1, dotted curves).

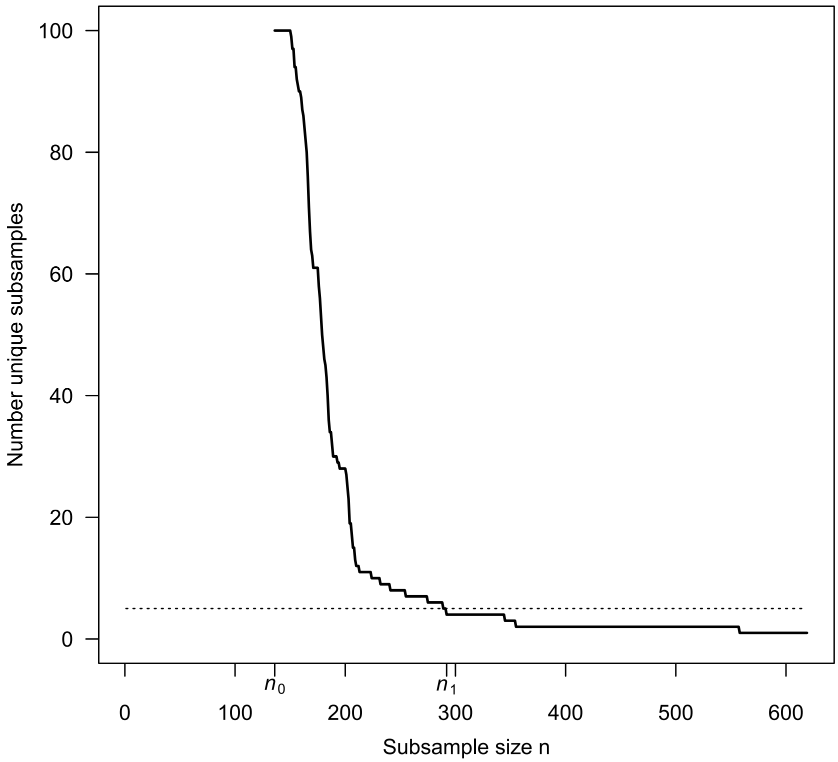


FIGURE 2 Number of unique subsamples across $B = 100$ forward searches going from $n_0 = 136$ to $N = 619$.

These curves decrease sharply between $n = 500$ and $n = 600$, which is where the cluster entered the subsample and the plots begin to differ markedly. After having entered the subsample, the residuals of the observations in the cluster show a sudden decrease followed by a gradual increase. As they entered the subsample, these observations affected the computation of $\overline{X}_j|X_+$ and, as a result, they came to resemble the conditional item means more, which then resulted in much smaller residuals. As more observations entered the subsample, the effect of the observations in the cluster on the computation of $\overline{X}_j|X_+$ decreased, and their residuals increased again. Also, the residuals of some other observations increased after the cluster entered the subsample and, as a result, these observations deviated more from $\overline{X}_j|X_+$ than previously.

Identifying influential observations. The gap plot (Figure 4) shows a large downward peak at $n = 544$, suggesting that a cluster of influential observations

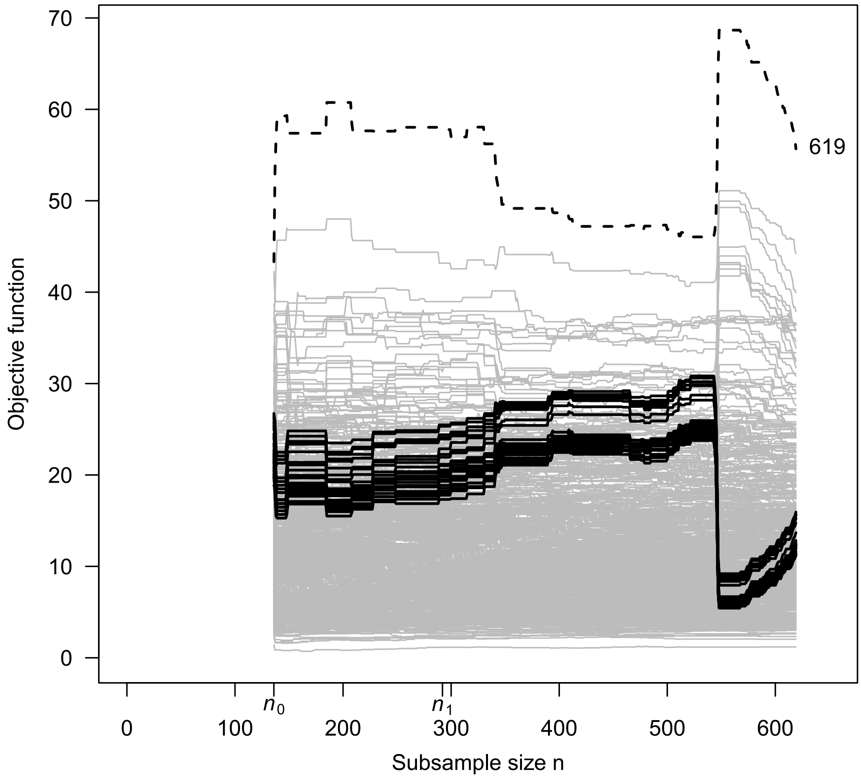


FIGURE 3 Forward plot of the objective function e_v^2. Thick dashed curve shows residual of suspect contaminant observation (#619) and solid curves show residuals of 30 influential contaminant observations (#589, ..., #618).

entered the subsample. Indeed, this cluster contained the 30 contaminants. Large negative values just after n_0 are due to the random selection of the initial subsample and should not be interpreted. Figure 5 shows follow-up plots for identifying the influential observations. Figure 5a shows follow-up plot-543 when observation #138 (thick black curve) entered the subsample. The vertical line represents the step when the observation(s) of interest entered the subsample. The thin black curves in all panels of Figure 5 are the 2.5%, 25%, 50%, 75%, and 97.5% percentile residuals based on the subsample at $n = 543$, which indicate what pattern of residuals may be expected. Observation #138 shows a pattern of residuals, which were large but consistent with the pattern of the other observations in the subsample and which cannot be considered influential. Figure 5b shows follow-up plot-544 when observation #597 entered the subsample (thick

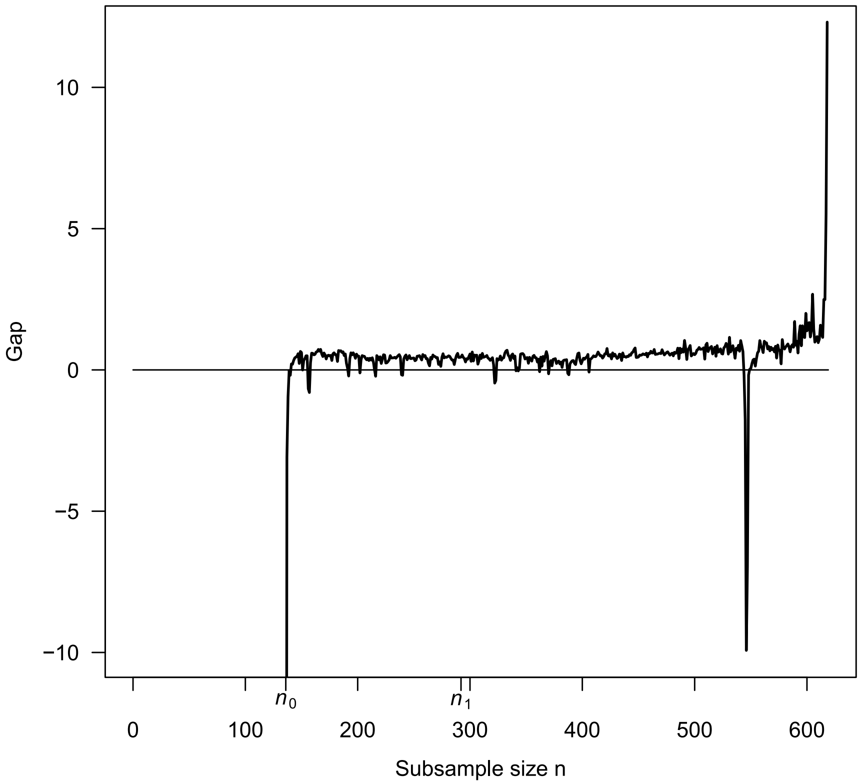
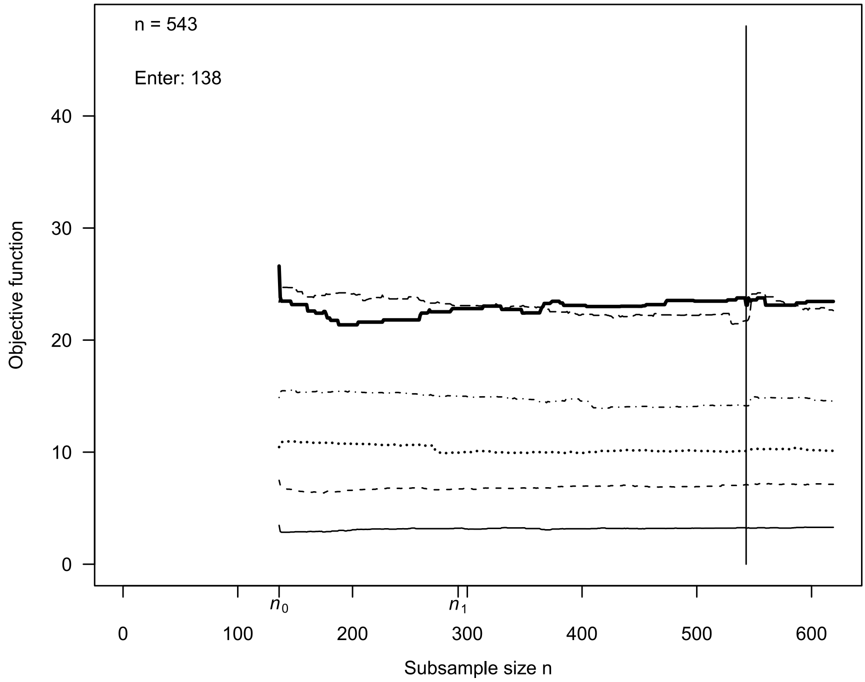


FIGURE 4 Gap plot.

black curve). This was the first influential contaminant observation that entered the subsample. Observation #597 was identified as an influential observation because the curve shows a sharp decrease (cf. Figure 1 dotted curve). The pattern of residuals of observation #597 across steps is markedly different from the pattern of the other 543 observations in the subsample.

Follow-up plots at $n = 545$ and $n = 546$ (not shown) are similar to Figure 5b. During these steps 11 influential observations were identified. Follow-up plot-547 (Figure 5c) shows that 18 observations entered the subsample (thick solid curves), all of which were identified as influential, and that 17 observations left the subsample (thick dotted curves). The residuals of some of these 17 observations were large from $n = 547$ onward. Because a large group of 18 influential observations entered the subsample, $\overline{X}|X_+$ changed dramatically at $n = 547$, thus causing these large residuals. Follow-up plot-548 (Figure 5d) shows that 9 observations entered the subsample. All showed a flat



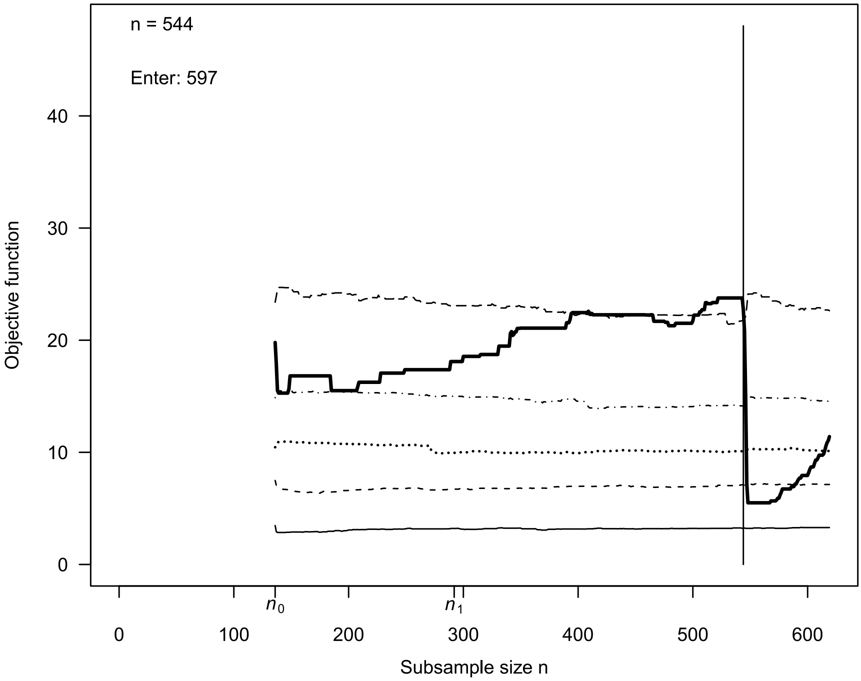
(a)

FIGURE 5 Forward plot of the observations entering and leaving the subsample at Steps 543 (a), 544 (b), 547 (c), and 548 (d). (*continued*)

pattern of residuals across steps (thick black curves) and were not considered influential. The 30 identified influential observations were the artificial influential contaminant observations.

Due to the presence of influential observations in the subsample, the forward search cannot be trusted to identify suspect observations. Hence, the influential observations were removed, and the forward search was rerun without the 30 influential observations but including the artificial suspect observation. We found that $n_0 = 119$ and $n_1 = 296$. No additional influential observations were found.

Finding n_2: Identifying suspect observations. The curves in Figure 6 show the *minexcl*-plot (thick curve) and Tukey's upper fence of the residuals of the observations in the subsample (thin curve). Tukey's upper fence intersected with the *minexcl*-plot between Steps $n = 556$ and $n = 557$; hence, $n_2 = 557$.

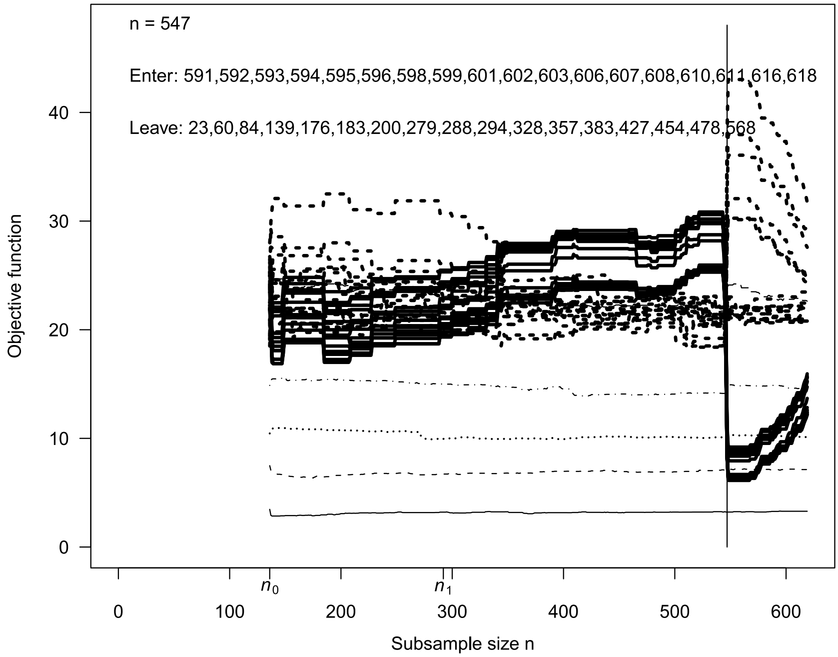


(b)

FIGURE 5 (*Continued*).

Thus, 32 suspect observations were identified. The plot was uneventful until subsample size n_2. After n_2, suspect observations entered the subsample showing a large increase in *minexcl*.

Inspecting the scale structure. The curve in Figure 7 shows that the number of scales found by the AISP changed across subsamples. From n_1 until $n = 430$, the 17 items usually formed a single scale; after that 2, 3, or 4 scales were found. One scale was longest throughout the forward search. The item-entry plot of this scale (Figure 8) shows that Items 2, 5, 6, 10, 12, and 15 belonged to this scale throughout the forward search and can be considered the core of the ACS. Items 1, 4, 9, 14, and 17 are borderline cases; they left the scale as more deviant and suspect observations entered the subsample. If the suspect observations were removed, these items would be included in the longest scale for most of the forward search. Items 3, 7, 8, 11, 13, and 16, on the other hand, left the longest scale rather early in the forward search and were candidates for removal.



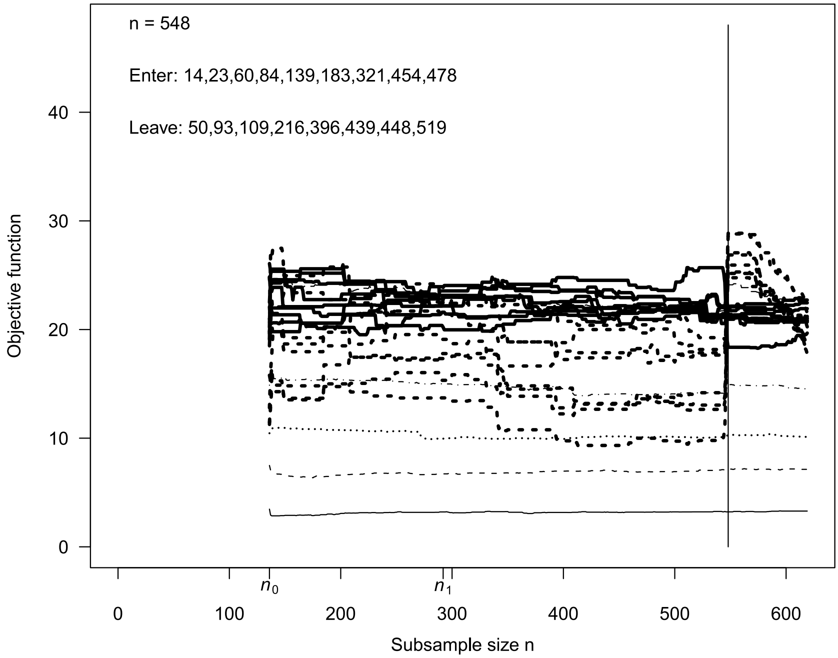
(c)

FIGURE 5 (*Continued*).

Results of the Second Forward Search Analysis

We wanted to retain as many items as possible; therefore, we only removed the items that were clearly not in the longest scale. The suspect observations remained in the data to investigate whether they affected the new scale. The second forward search was run using the 11 core items and borderline items (1, 2, 4, 5, 6, 9, 10, 12, 14, 15, and 17) and $N = 589$ respondents. We found that $n_0 = 77$, $n_1 = 302$, and $n_2 = 558$. The forward plot of the objective function and the gap plot (not shown) were uneventful. No additional influential observations were detected between n_1 and n_2.

Estimated IRF: Identifying suspect influential observations. Due to limited space, only the forward search of the estimated IRF of Item 1 is discussed extensively, and only highlights of the forward search of the estimated IRFs for the other items are mentioned. Figure 9 shows the conventional display of the item-restscore regression for item 1, $\overline{X}_1|R_{(-1)}$, for the entire sample (solid lines,



(d)

FIGURE 5 (*Continued*).

open circles) and the subsample at $n = 302$ (dashed lines, solid circles). The plot was based on seven restscore groups ($G_1 = 7$) yielding seven estimated points of the IRF, connected by straight lines.

Figure 10 shows the forward plot of the estimated IRF, $\overline{X}_1|R_{(-1)}$. Each restscore group is represented by a separate curve. The seven open circles at the end of the forward search correspond to the seven open circles in Figure 9, and the seven solid circles at $n = 302$ correspond to the seven solid circles in Figure 9. If manifest monotonicity (Junker & Sijtsma, 2000) holds for an estimated IRF (cf. Figure 9, open circles), it is expected that between n_1 and N the curves in Figure 10 are ordered without intersections from the joined restscore groups $R_{(-1)} \in \{31 - 39\}$ (highest) to $R_{(-1)} \in \{9 - 18\}$ (lowest). This was found to be true for the majority of the subsamples, yet small intersections were found for restscore Groups 2 and 3 and restscore Groups 4 and 5 (Figure 10). The curves of all restscore groups showed small changes in $\overline{X}_1|R_{(-1)}$ when moving forward in the plot. The forward plots of the other items showed similar patterns for the curves of the restscore groups. The 32

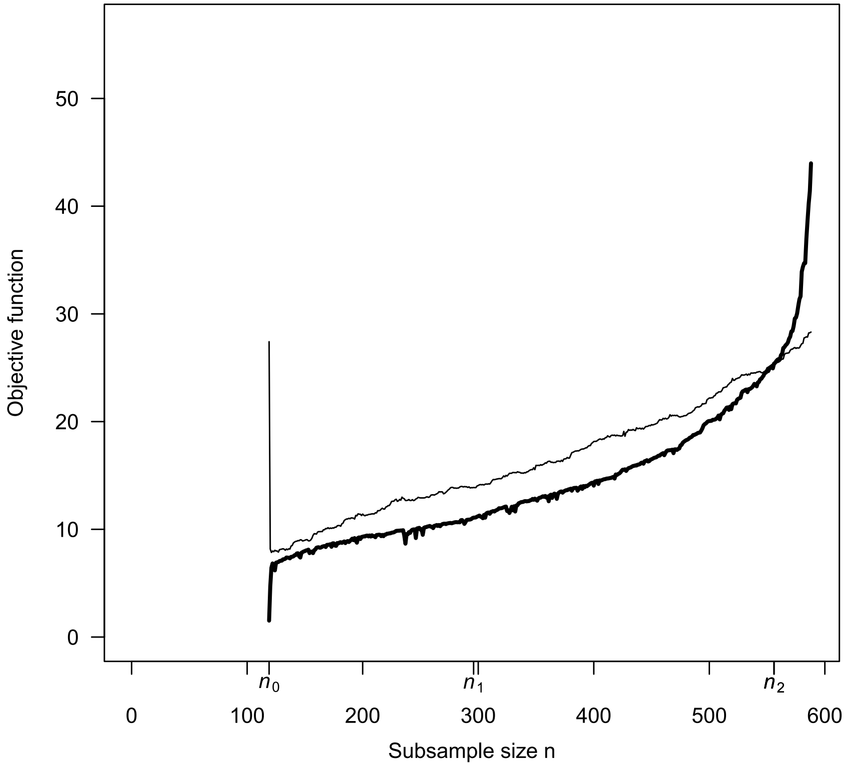


FIGURE 6 *Minexcl*-plot (thick curve) and Tukey's upper fence for the residuals in the subsample (thin curve).

suspect observations, which entered the subsample after n_2, did not affect the estimated IRFs.

Scalability coefficients: Identifying suspect influential observations. Figure 11 shows the forward plot of the scalability coefficients H (thick curve) and H_j (thin curves), and it also shows the default lower bound value $c = .3$ (horizontal straight line). As expected, scalability decreased with increasing subsample size. Curves of the scalability coefficients were uneventful between n_1 and N suggesting that the suspect observations did not seem to exercise a significant influence on the scalability coefficients. However, if the default lower bound $c = .3$ is used as a criterion, the scaling results were rather different with and without the suspect observations. With suspect observations, the items formed a weak scale $H = .30$ but $H_j < c$ for items 1, 2, 4, 9, 14, and 17;

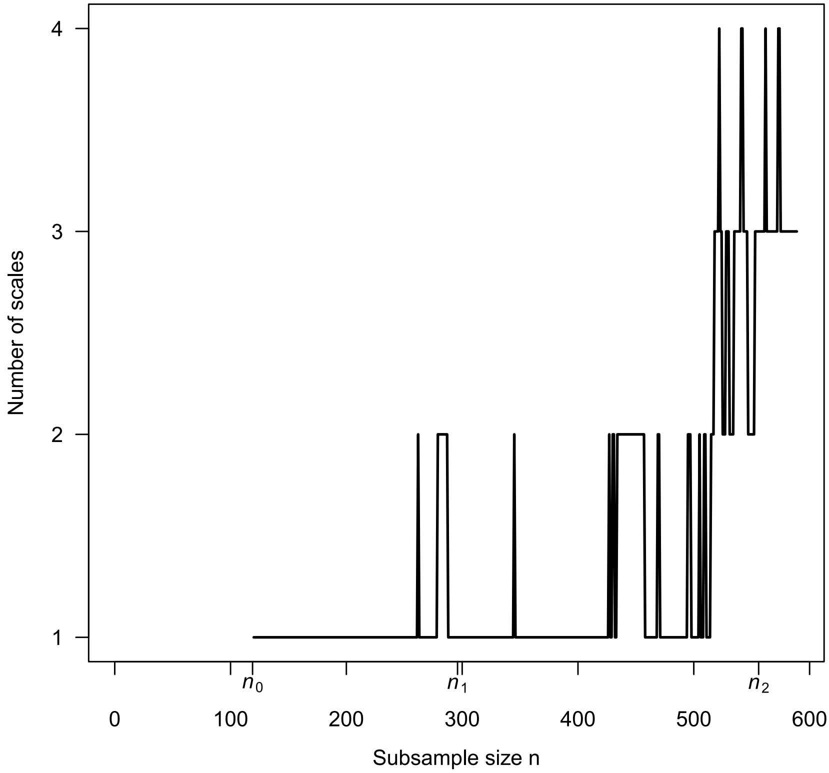


FIGURE 7 Number of scales found by the automated item selection procedure (AISP).

without suspect observations, the items also formed a weak scale ($H = .34$), but only two item scalability coefficients were slightly lower than .3 (i.e., $H_1 = .29$ and $H_{17} = .29$). Appendix A shows all H_j values.

What if Influential Observations Had Not Been Removed?

The effect of the influential observations on the partitioning of the item set into scales was small. The item-entry plot (not shown) was more difficult to interpret than Figure 8 but we would have chosen the same 11 items for the second forward search. Figure 12a shows the estimated IRFs (cf. Figure 10) and Figure 12b shows the scalability coefficients (cf. Figure 11) in case the influential observations were included. The estimated IRF of restscore Group 2 (containing the influential observations) changed dramatically at step $n = 547$; the other estimated IRFs were not affected. The influential observations

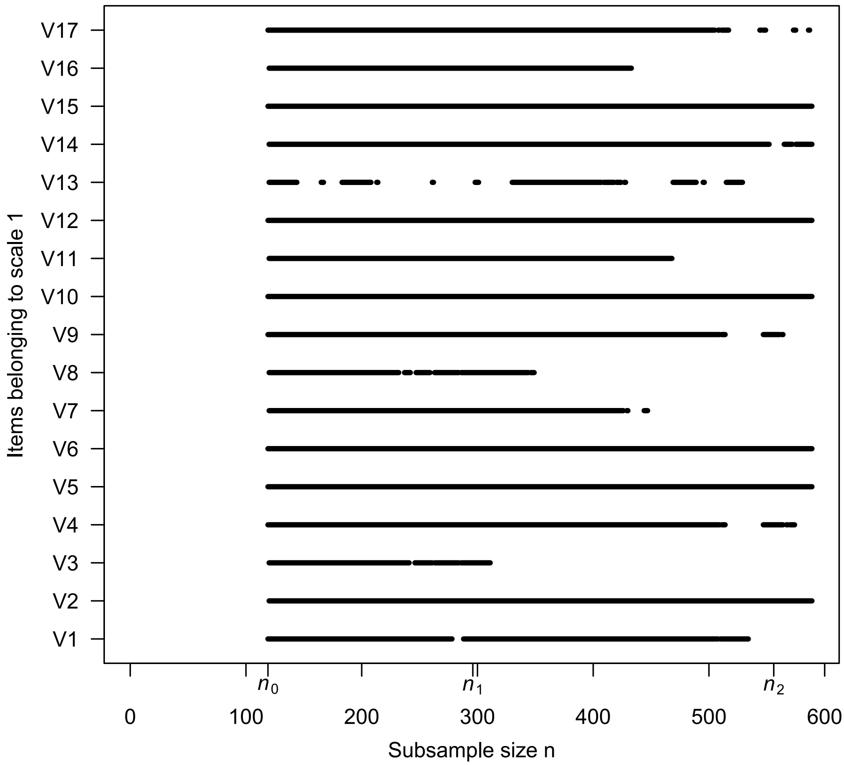


FIGURE 8 Item-entry plot for the longest scale in the automated item selection procedure (AISP).

also had a decreasing effect on H and on the item scalability coefficients of 3 items (6, 9, and 17) showing a huge decrease at $n = 547$ (H_js shown in Appendix A). These plots indicate that the influentials indeed affected the results of MSA.

DISCUSSION

The adaptation of the forward search to MSA showed promising results. Based on the forward plots, suspect observations and influential observations could be identified, which might otherwise remain undetected. The forward plots provide the researcher with diagnostic information about observations that are different from the remainder of the data set or have a large influence on the statistical analysis. The key feature of the forward search is that it very quickly

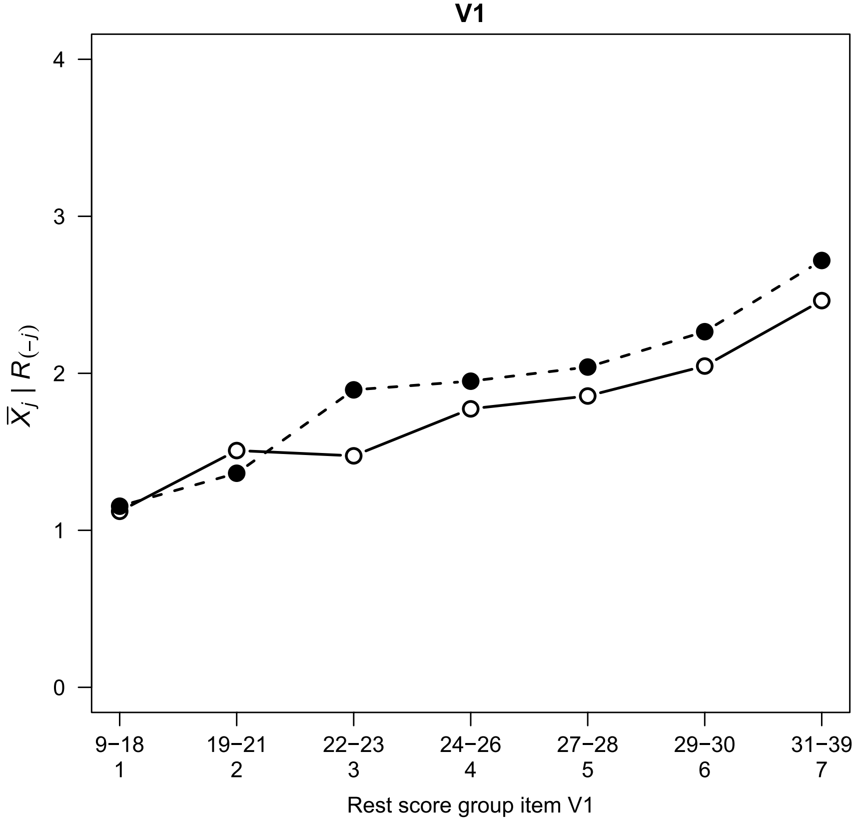


FIGURE 9 Estimated item response function (IRF) of Item 1 for $n = 302$ in open circles and $n = 589$ in solid circles. Estimates were based on seven combined restscore groups. The horizontal axis showes the range and the number of the combined restscore groups.

identifies a subsample of normal observations, which serves as a benchmark for unselected observations that may be suspect of influential. Traditional outlier detection methods assess individual observations amidst the complete sample of observations including possible suspect and influential observations. Obviously, this may contaminate conclusions about individual observations.

Two aspects of the forward search could not easily be adapted to MSA. First, residuals can be standardized in the traditional forward search (Atkinson & Riani, 2004, pp. 66–67) but not in the adapted forward search. Standardizing residuals does not affect the traditional forward search but facilitates the interpretation of the residuals. For the adapted forward search, residuals can be standardized in at least two ways: standardization of residuals for each item separately and

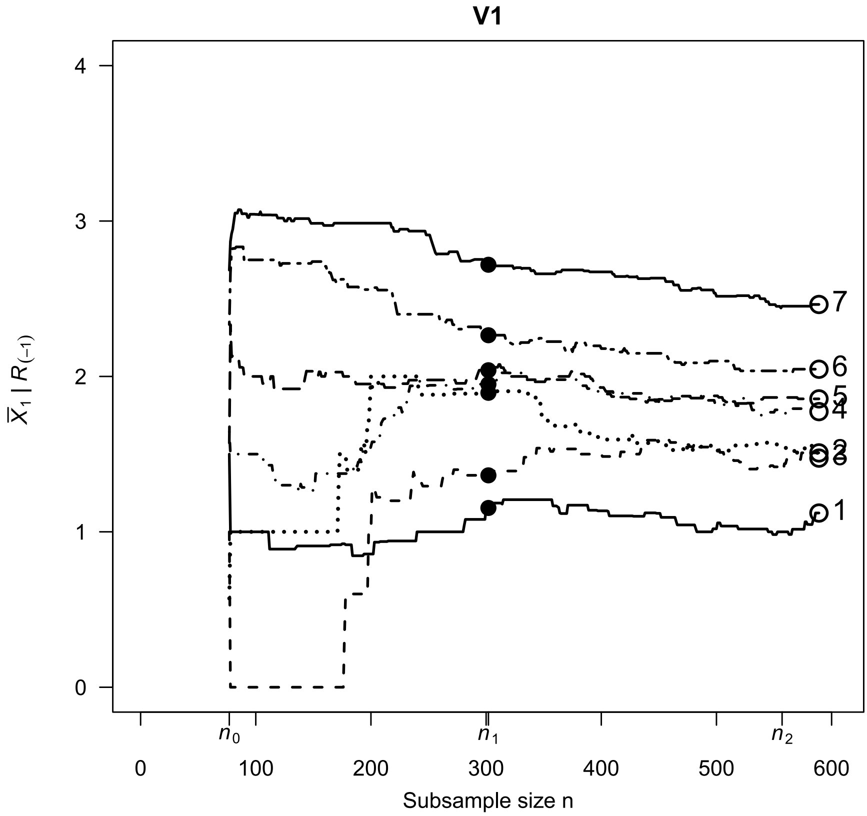


FIGURE 10 Forward plot of the estimated item response function (IRF) $\overline{X}_1|R_{(-1)}$. There were seven combined restscore groups: $1 = \{9-18\}$, $2 = \{19-21\}$, $3 = \{22-23\}$, $4 = \{24-26\}$, $5 = \{27-28\}$, $6 = \{29-30\}$, and $7 = \{31-39\}$.

standardization of residuals over items. The former method has the disadvantage that it affects the results of the forward search, and the latter method has the disadvantage that it is not clear how a pooled standard error over items should be computed. Second, in the *minexcl*-plot, the traditional forward search provides confidence envelopes, but we could not provide these envelopes for the adapted forward search. The core of the problem is that the construction of confidence envelopes requires a parameterization of the underlying model, but the model underlying MSA, which is the MHM, is a nonparametric model that leaves much freedom to the item response functions as long as they do not decrease. This freedom hampers the bootstrap necessary for obtaining the confidence envelopes. Instead, for the *minexcl*-plot we computed Tukey's fences (Figure 6), which serve

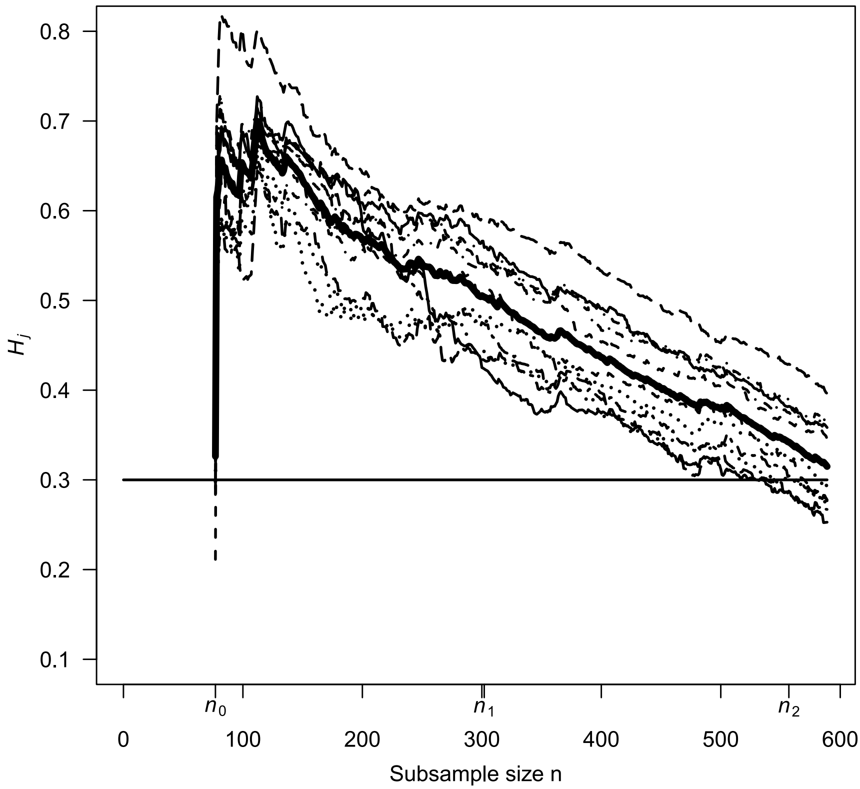
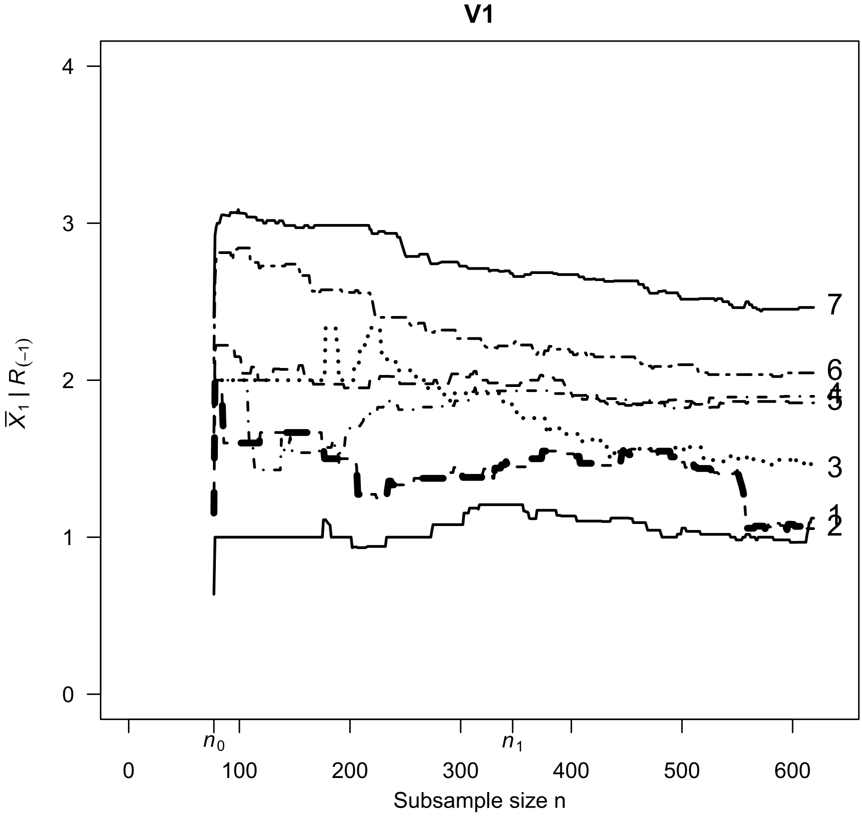


FIGURE 11 Forward plot of the item scalability coefficients H_j (thin curves) and the total-scale H (thick curve).

the same purpose as the confidence envelopes. Both issues are topics for further research.

The identification of suspect observations and influential observations is the first step in outlier analysis. We recommend running the final data analyses with and without the identified influential and suspect observations. If the outcomes differ markedly, the results from the data analysis without the influentials and suspects should be trusted and reported. The identification of a cluster of influential observations may suggest that the sample contains data from multiple populations. The different populations may be treated separately or an alternative model may be defined that takes the different populations into account.

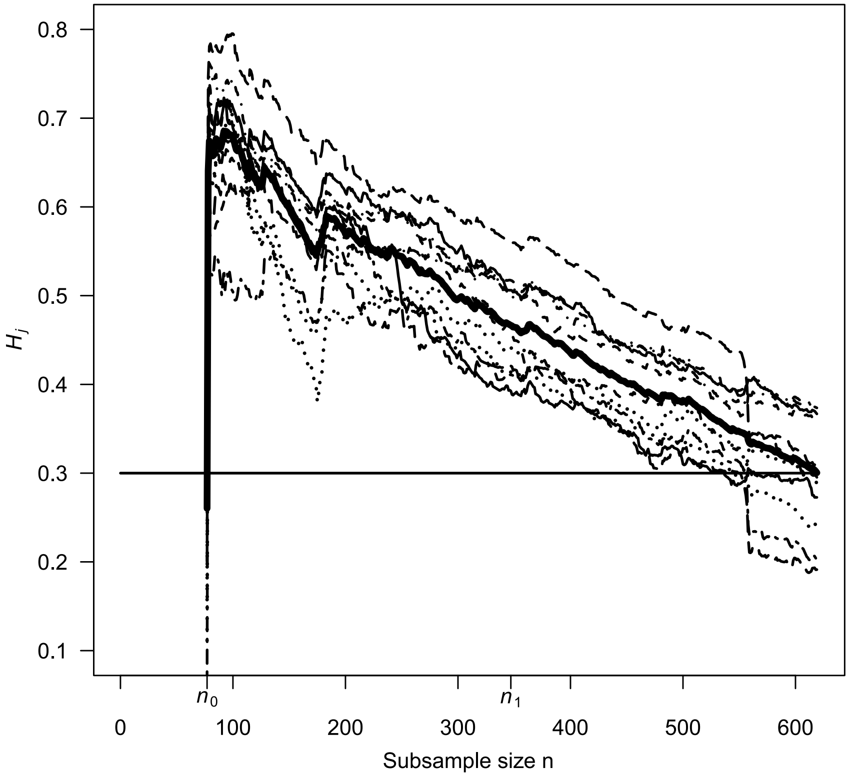
The adaptation of the forward search to MSA entailed defining a suitable objective function and a procedure for selecting the set of items that could serve



(a)

FIGURE 12 Forward plot of the estimated item response function (IRF) $\overline{X}_1|R_{(-1)}$ (panel a) and forward plot of the item scalability coefficients H_j (thin curves) and the total-scale H (thick curve) (panel b) for the contaminated data set. (*continued*)

as the basis of the forward search. Alternative functions and item selection procedures may be considered, but given our experience with MSA we are convinced that the decision to focus on item-score vectors and evaluate their consistency with respect to the set of IRFs is a solid one. Not only do item-score vectors represent the basic individual data in an MSA with respect to the IRFs but also item-score vectors lie at the basis of computing scalability coefficients, which are another key asset in an MSA. Thus, identifying suspect and influential observations on the basis of item-score vectors relative to IRFs is an excellent starting point for an MSA data analysis.



(b)

FIGURE 12 (*Continued*).

REFERENCES

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association, 89*, 1329–1339.

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York, NY: Springer.

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis, 52*, 272–285.

Atkinson, A. C., Riani, M., & Cerioli A. (2004). *Exploring multivariate data with the forward search*. New York, NY: Springer.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. New York, NY: Wiley.

Bech, P., Wilson, P., Wessel, T., Lunde, M., & Fava, M. (2009) A validation analysis of two self-reported HAM-D-6 versions. *Acta Psychiatrica Scandinavica, 119*, 298–303.

Beckman, R. J., & Cook, R. D. (1983). Outlier..........s. *Technometrics, 25*, 119–149.

Bekker, M. H. J., & Van Assen, M. A. L. M. (2006). A short form of the autonomy scale: Properties of the autonomy-connectedness scale (ACS-30). *Journal of Personality Assessment, 86*, 51–60.

Bertaccini, B., & Varriale, R. (2007). Robust analysis of variance: An approach based on the forward search. *Computational Statistics & Data Analysis, 51*, 5172–5183.

Chambers, R., Hentges, A., & Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A, 167*, 323–339.

Cheng, T.-C., & Biswas, A. (2008). Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data. *Computational Statistics & Data Analysis, 52*, 2042–2065.

Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B, 54*, 761–771.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychology Measurement, 19,* 337–352.

Jacoby, W. G. (2008). Comment: The dimensionality of public attitudes toward government spending. *Political Research Quarterly, 61*, 158–161.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65–81.

Korner, A., Lauritzen, L., Abelskov, K., Gulmann, N. C., Brodersen, A. M., Wedervang-Jensen, T., & Kjeldgaard, K. M. (2007). Rating scales for depression in the elderly: External and internal validity. *Journal of Clinical Psychiatry, 68*, 384–389.

Loner, E. (2008). The importance of having a different opinion: Europeans and GM foods. *European Journal of Sociology, 49*, 31–63.

Mavridis, D., & Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research, 43*, 453–475.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin, Germany: De Gruyter.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows:. User's manual.* Groningen, The Netherlands: iec ProGAMMA.

Paas, L. J., & Sijtsma, K. (2008). Nonparametric item response theory for investigating dimensionality of marketing scales: A SERVQUAL application. *Marketing Letters, 19*, 157–170.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection.* New York, NY: Wiley.

Santtila, P., Pakkanen, T., Zappala, A., Bosco, D., Valkama, M., & Mokros, A. (2008). Behavioural crime linking in serial homicide. *Psychology Crime & Law, 14*, 245–265.

Sijtsma, K., Meijer, R. R., & Van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences, 50*, 31–37.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*, 575–579.

Watson, R., Roberts, B., Gow, A., & Deary, I. J. (2008). A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences, 45*, 333–335.

Zijlstra, W. P. (2010). fwdmsa: An R package for the forward search for Mokken scale analysis, version 0.1. Available from http://cran.r-project.org/web/packages/fwdmsa/index.html

Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*, 531–555.

Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2011). Outliers in test and questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*. doi: 10.3102/1076998610366263

APPENDIX A

The 17 items from the Dutch version of the subscale "sensitivity to others" of the Autonomy-Connectedness Scale (Bekker & Van Assen, 2006). Each item has five ordered answer categories (0 = *disagree*, 1 = *slightly disagree*, 2 = *agree nor disagree*, 3 = *slightly agree*, 4 = *agree*). An asterisk (∗) indicates a negatively worded item. For the 11 items in the second forward search analysis, the scalability coefficients H_j are shown (∗∗: influential and suspect observations included; †: influential observations removed, suspects included; ‡: influential and suspect observations removed).

Item	*Label*	H_j^{**}	$H_j^{\dagger}$	$H_j^{\ddagger}$
∗1.	Usually I can dismiss another person's misery from my mind	.27	.25	.29
2.	If I have things my own way against the will of others, I usually get very restless	.30	.28	.32
3.	I hate detachment			
∗4.	I am seldom occupied with the feelings and experiences of others	.24	.29	.33
∗5.	I easily put aside other people's comments	.37	.36	.38
∗6.	I am rarely occupied with other people's view of me	.31	.40	.43
7.	If I imagine myself having to say good-bye to a beloved person, I feel brokenhearted in advance			
∗8.	I am seldom inclined to ask other people's advice			
9.	I often go deeply into other people's feelings	.20	.28	.30
10.	I often wonder what other people think of me	.37	.36	.38
∗11.	When I make important decisions about my life, I leave other people's wishes and opinions out of consideration			
12.	I can hardly bear it when other people are angry with me	.36	.35	.37
13.	Somebody else's experiences leave a strong mark on my own moods			
14.	I feel a strong need for other people's advice and guidance	.29	.27	.30
∗15.	If I do something that bothers other people, I can easily dismiss that from my mind	.37	.36	.38
16.	I often long for love and warmth			
∗17.	I can easily back out of things that people who are important to me want me to do	.19	.27	.29

APPENDIX B

```
# First Forward Search Analysis
  library(fwdmsa)
  data(acs.cont)
# Determining n1 = 292
# Takes approximately 40 minutes
  fs1.1.n1 <- fs.MSA.n1(acs.cont, B=100)
  n1 <- fs1.1.n1$n1
```

```
# Figure 2: Plot of number unique subsamples
  plot(fs1.1.n1)
# Running the forward search
  fs1.1 <- fs.MSA(acs.cont)
# Figure 3: Plot of objective function
  plot(fs1.1, col="gray70", n0=TRUE, n1=fs.res.cont.n1$n1, xlim=c(0,650))
  plot(fs1.1, id.observation=619, col=1, lwd=2, lty=2, add=TRUE)
  plot(fs1.1, observations=589:618, lwd=2, add=TRUE)
# Figure 4: Gap plot
  plot(fs1.1, type="gap", ylim=c(-10,12), n0=TRUE, n1=292)
# Figure 5: Follow-up plots
  plot(fs1.1, type="followup", step=543:548, reference.step=543, n0=TRUE, n1=292)

# Remove influential observations from the data
  acs.sus <- acs.cont[-(589:618),]
# Determining n1 = 296
  fs1.2.n1 <- fs.MSA.n1(acs.sus, B=100)
  n1 <- fs1.2.n1$n1
# Running the forward search
  fs1.2 <- fs.MSA(acs.sus)
# Figure 6: Minexcl plot
  plot(fs1.2, type="minexcl", n0=TRUE, n1=296, n2=TRUE)
# Figure 7: Plot of number of scales
  plot(fs1.2, type="num.scale", n0=TRUE, n1=296, n2=TRUE)
# Figure 8: Item=entry plot for the longest scale
  plot(fs1.2, type="scale", id.scale=1, n0=TRUE, n1=296, n2=TRUE)

# Second Forward Search Analysis
# Remove bad items from the data
  acs.min.core <- acs.cont[-(589:618),-c(3,7,8,11,13,16)]
# Determining n1= 302
  fs2.n1 <- fs.MSA.n1(acs.min.core, B=100)
  n1 <- fs2.1.n1$n1
# Running the forward search
  fs2 <- fs.MSA(acs.min.core)
# Figure 9: Plot of restscore regression of item 1 for steps 302 and 589
  plot(fs2, type="restscore", step=302, items=1, lty=2, n0=TRUE, n1=302, n2=TRUE)
  plot(fs2, type="restscore", step=589, items=1, lty=1, add=TRUE)
# Figure 10: Plot of estimated IRF of item 1
  plot(fs2, type="IRF", items=1, n0=TRUE, n1=302, n2=TRUE)
# Figure 11: Plot of coefH
  plot(fs2, type="coefH", n0=TRUE, n1=302, n2=TRUE, ylim=c(.1,.8))

# What if influential observations were not removed from the data
  acs.cont.core <- acs.cont[,-c(3,7,8,11,13,16)]
# Determining n1 = 347
  fs3.n1 <- fs.MSA.n1(acs.cont.core, B=100)
  n1 <- fs3.n1$n1
# Running the forward search
  fs3 <- fs.MSA(acs.cont.core)
# Figure 12a: Plot of estimated IRF of item 1 with influential observations
  plot(fs3, type="IRF", items=1, n0=TRUE, n1=347, n2=FALSE)
# Figure 12b: Plot of coefH with influential observations
  plot(fs3, type="coefH", n0=TRUE, n1=347, n2=FALSE, ylim=c(.1,.8))
```