Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

A Model-Free Diagnostic for Single-Peakedness of Item Responses Using Ordered Conditional Means

Marike Polak , Mark de Rooij & Willem J. Heiser

To cite this article: Marike Polak , Mark de Rooij & Willem J. Heiser (2012) A Model-Free Diagnostic for Single-Peakedness of Item Responses Using Ordered Conditional Means, Multivariate Behavioral Research, 47:5, 743-770, DOI: 10.1080/00273171.2012.715563

To link to this article: https://doi.org/10.1080/00273171.2012.715563

Multivariate Behavioral Research, 47:743–770, 2012

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.715563

Psychology Press
Taylor & Francis Group

A Model-Free Diagnostic for Single-Peakedness of Item Responses Using Ordered Conditional Means

Marike Polak
Institute of Psychology, Erasmus University Rotterdam

Mark de Rooij and Willem J. Heiser
Psychological Institute, Leiden University

In this article we propose a model-free diagnostic for single-peakedness (unimodality) of item responses. Presuming a unidimensional unfolding scale and a given item ordering, we approximate item response functions of all items based on ordered conditional means (OCM). The proposed OCM methodology is based on Thurstone & Chave's (1929) *criterion of irrelevance*, which is a graphical, exploratory method for evaluating the "relevance" of dichotomous attitude items. We generalized this criterion to graded response items and quantified the relevance by fitting a unimodal smoother. The resulting goodness-of-fit was used to determine item fit and aggregated scale fit. Based on a simulation procedure, cutoff values were proposed for the measures of item fit. These cutoff values showed high power rates and acceptable Type I error rates. We present 2 applications of the OCM method. First, we apply the OCM method to personality data from the Developmental Profile; second, we analyze attitude data collected by Roberts and Laughlin (1996) concerning opinions of capital punishment.

In the field of psychometrics one major objective is to construct valid and reliable scales. Such scales are constructed to measure latent constructs, for example,

Correspondence concerning this article should be addressed to Marike Polak, Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Wassenaarseweg 52, P.O. Box 9555, Leiden, The Netherlands. E-mail: Polak@fsw.eur.nl

attitudes, and consist of a set of items, for example, numerical attitude items. The goal of item analysis is to determine whether there are one or more underlying scales and whether items need to be discarded or included. In this article we focus on scale and item evaluation for unidimensional unfolding scales consisting of a set of ordered polytomous items.

The essence of an unfolding scale is that the probability of agreement with a certain item on this scale is inversely related to the distance between the position of the item on the latent continuum and the position of the respondent; the closer an item is located near the respondent's position on the latent continuum, the more likely the respondent will agree with it. Because this yields a single-peaked response function, these items are often referred to as single-peaked items. Single-peaked items typically arise in fields of personality measurement (e.g., Chernyshenko, Stark, Drasgow, & Roberts, 2007; Weekers & Meijer, 2008), vocational interest assessment (e.g., Tay, Drasgow, Rounds, & Williams, 2009), preference research (e.g., Ashby & Ennis, 2002) and attitude research (e.g., Andrich & Styles, 1998; Carter & Dalal, 2010).

Historically, methods for item analysis and unfolding scale construction were developed by Thurstone (1927, 1928; Thurstone & Chave, 1929). More recently, a number of item response theory (IRT) models have been developed that explicitly take single-peakedness of the item response function (IRF) into account. Examples are the nonparametric multiple unidimensional unfolding model (MUDFOLD; Van Schuur, 1992), the parametric PARELLA model (Hoijtink, 1991), the hyperbolic cosine model (Andrich & Luo, 1993), the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), the multidimensional normal PDF model (Maydeu-Olivares, Hernandez, & McDonald, 2006), and the multidimensional unfolding model (Javaras & Ripley, 2007). These models are often referred to as unfolding IRT models (see Andrich, 1996, for an introduction to this type of models). They differ in the exact set of assumptions or restrictions they employ, including whether they handle items with more than two response categories or whether the latent space is allowed to be multidimensional.

Within the framework of unfolding IRT, tools for item analysis include a modest number of item fit diagnostics. The most well known are the item fit statistics and diagnostics available in the GGUM software, GGUM2004 (Roberts, Fang, Ciu, & Wang, 2006). At present, Roberts (2008) recommends the likelihood-based indices of fit that have been generalized to polytomous unfolding models from Orlando and Thissen (2000). Another contribution to determining item fit for parametric dichotomous unfolding IRT models were two Lagrange Multiplier tests for differential item functioning and violation of the shape of the IRFs by Weekers (2009). A measure to assess unidimensionality for unfolding IRT models is the Q_3 statistic (Habing, Finch, & Roberts, 2005) that was generalized from Yen (1984, 1993) to single-peaked items. A diagnostic

based on nonparametric IRT that is available in MUDFOLD 4.0 (Van Schuur & Post, 1998) is the conditional adjacency matrix (CAM; Post, 1992).

Unfolding IRT has brought substantive advances to applied fields, such as personality measurement, by providing theoretically more sound and technically more flexible models compared with traditional psychometric methods (Chernyshenko et al., 2007). However, the added complexity of these models also brings some drawbacks, especially for practical application.

One drawback is that the item diagnostics provided by unfolding IRT models are conditional on the model estimates. Model violations, such as misfit of the IRF, violation of unidimensionality,[1] or violation of the "ideal point process" assumption (i.e., when data are dominance data) may result in nonconvergence of parameter estimation (cf. Tay, Ali, Drasgow, & Williams, 2011). Nonconvergence is considered a signal of serious misfit, but when this happens, the practical researcher lacks information about possible causes for this misfit. Item misfit may or may not be the cause of the estimation problems. Software such as GGUM2004 and MUDFOLD 4.0 may for this reason require a preselection of items. This preselection is not yet straightforward. The GGUM2004 manual recommends the use of principal component analysis (PCA), which is not optimal for analyzing data conforming to an unfolding model (cf. Van Schuur & Kiers, 1994) and thus introduces other difficulties.

Related to the problem of nonconvergence is the requirement of relatively large samples (i.e., $N > 1{,}000$) to obtain accurate and stable parameter estimates, and consequently, accurate item fit statistics for unfolding IRT models.

Moreover, discarding an item based on a significant fit statistic alone is still not recommended for GGUM2004 because Type I error rates have been shown to depend on factors such as sample size, test length, or number of response categories (DeMars, 2004; Roberts, 2008; Roberts & Thompson, 2011). The GGUM2004 technical manual therefore recommends the use of graphical methods to investigate IRFs as an extra check.

The main objective of this article is to develop a model-free methodology for scale and item evaluation of single-peaked items that can be used prior to, or in combination with, any of the existing unfolding (IRT) models. The proposed ordered conditional means (OCM) method gives a graphical display of the approximated IRF for each item. Based on the observed scores and a given item order, ordered conditional means are depicted in so-called OCM diagrams. The method assumes unidimensionality, an ordering of the items along

[1] Carter and Zickar (2011) showed that for GGUM simulated data nonconvergence of the GGUM estimation was relatively rare, although under several conditions of bidimensionality item parameters were degenerate (too extreme). However, it should be noted that in those instances of convergence of the estimation procedure with degenerated parameter estimates, the GGUM item fit statistics were able to detect the deviant items.

the underlying dimension, and single-peaked IRFs but does not impose any further constraints on the shape of the IRF. A smoother is used as nonparametric fitting procedure, which results in values of fit for all items and the scale as a whole.

The proposed OCM approach and the existing IRT approaches might be used at different stages of the process of scale construction. For instance, an adequate item set may be selected first by the OCM methodology, and in a further step, an unfolding IRT model may be fitted to obtain more precise information about the item functioning.

Altogether, the proposed methodology provides answers to questions regarding the practical use of unfolding scales that were discussed by Dalal, Withrow, Gibby, and Zickar (2010). First, we discuss how unfolding measures can be scored without IRT-based parameters. Second, by discussing the OCM methodology, we show how an unfolding scale can be evaluated based on a relatively small sample.

This article is organized as follows: The next section presents some approaches to obtain an ordering for a set of single-peaked items. The model-free OCM methodology for scale and item evaluation of single-peaked items is explained in the subsequent section. In the fourth section, a practical diagnostic procedure is developed. Subsequently, we present a simulation study and two real data examples. A discussion is given in the final section, and technical details are given in the two appendices.

ORDERING SINGLE-PEAKED ITEMS

Typical for unidimensional unfolding scales is that items cover the entire (bipolar) latent continuum. Estimating item and person locations on the latent continuum is an explicit goal of unfolding scale analysis. This is fundamentally different from the Likert procedure, where only extreme items are selected and individual differences are computed based on total scores or proportion endorsed. A respondent's position on the unfolding scale cannot be derived from his total score. Rather, his position is determined by computing the mean position associated with endorsed items.

Thurstone (1928) was the first who showed that once the item order on the latent scale is known, an approximate IRF could be derived. This principle is elaborated on in the subsequent section. The OCM methodology we propose in this article also builds on the principle that an approximate IRF can be derived once the item ordering is known. In this section we briefly discuss methods that, besides the unfolding IRT models that were listed in the introduction, can be used to obtain an objective ordering of the data. In particular we propose correspondence analysis (CA) for this purpose. Note that the scope of this article

is restricted to binary or graded item response data organized in a rectangular matrix.

Procedures for ordering single-peaked items are also known as seriation methods. Seriation methods find an optimal ordering of the rows and column of the data matrix so that data show a "simplex-like" pattern, also referred to as Robinson pattern (cf. Hubert, Arabie, & Meulman, 1998). That is, when the items and persons are ordered according to their location on the latent scale, the scores along the diagonal of the matrix will be high, whereas moving downward and to the lower left-hand corner, the scores will decrease to zero.

A description of the development of seriation in the field of archeology is given by Ihm (2005). Ter Braak and Prentice (1988, 2004) review data analysis techniques for seriation (or ordination) in ecology. In both fields CA has become a popular seriation technique. The use of CA for seriation in psychometrics is described in Gifi (1990); Heiser (1981); and recently in Polak, Heiser, and De Rooij (2009).

CA, which is related to PCA, is available in SPSS (Categories Module; Meulman & Heiser, 2004); in SAS/STAT (CORRESP procedure; SAS Institute, 2008); or in R, for instance, in the `anacor` package (De Leeuw & Mair, 2009) and the `vegan` package (Oksanen et al., 2010). See Appendix A for mathematical details and a brief theoretical explanation of CA.

Besides CA, we know of two ordinal methods for scaling single-peaked items. First, Cliff, Collins, Zatkin, Gallipeau, and McCormick (1988) presented a nonparametric and deterministic method for ordering the rows and columns of a rank-order matrix in order to obtain as close a Robinson pattern as possible. Second, Johnson (2006) provided an algorithm that uses the CAM by Post (1992) to estimate the rank order of items and respondents on the latent scale.

MODEL-FREE METHOD FOR ITEM EVALUATION: ORDERED CONDITIONAL MEANS (OCM)

The proposed methodology builds on Thurstone and Chave's (1929) *criterion of irrelevance* (COI), which is a classic exploratory method for evaluation of single-peaked dichotomous attitude items. The COI is based on a diagram for a certain item g, depicting on the horizontal axis the position of all items on the scale and on the vertical axis the index of similarity (C_{gh}) of item g with any other item h on the scale. The index C_{gh} is defined as

$$C_{gh} = \frac{N(g,h)}{N(h)} \tag{1}$$

where $N(g,h)$ is the number of participants in the sample choosing both items g and h, and $N(h)$ is the number of participants in the sample choosing item

h. Thus, C_{gh} is the conditional probability of choosing item g given that a participant chooses item h.

Thurstone and Chave's (1929) method for determining item fit was based on the visual inspection of the COI diagrams. The key idea was that the more the diagram of item g showed a single-peaked pattern, the more "relevant" item g was for discriminating between different attitudes.

In this section we propose a generalization of the COI that is suited for binary as well as polytomous items. In this article, we define the polytomous extension of Thurstone and Chave's (1929) index C_{gh} as

$$C_{gh} = \frac{1}{M} E(Z_{ig} | Z_{ih} = M), \tag{2}$$

which is estimated by

$$\hat{c}_{gh} = \frac{1}{M} \frac{1}{N_h} \sum_{i \epsilon L_h} Z_{ig}, \tag{3}$$

where Z_{ig} is the observed response of participant i $(i = 1, \ldots, n)$ to item g $(g = 1, \ldots, k)$ with $Z_{ig} = z$, where $z = 0, 1, \ldots, M$, where $z = 0$ indicates the strongest level of disagreement, and $z = M$ indicates the strongest level of agreement. L_h is the subset of participants in the sample with $Z_{ih} = M$ and N_h is the number of participants in subset L_h.

Thus, $\hat{c}_{gh}$ equals the estimated conditional mean response on item g for those participants who express the highest level of agreement with item h, where $\hat{c}_{gh}$ is scaled within the 0−1 interval by rescaling to obtain 1 as a maximum and 0 as a minimum. In case of binary responses, Equation (3) equals Equation (1).

The proposed OCM diagram for a certain item g depicts on the horizontal axis the rank numbers of all items on the scale and on the vertical axis the generalized index of similarity of item g with any other item h, as defined in Equation (3). Analogous to Thurstone and Chave (1929), we regard the OCM diagram of item g as a representation of the IRF of item g. We explain the OCM diagrams with the following example.

Suppose we have a simulated data set of 300 persons rating nine items, $v1$ to $v9$, using a 5-point scale ranging from 0 to 4. Response probabilities were generated according to the GGUM (see Appendix B for the GGUM formula and a brief description). Item locations (δ_g) were equally spaced ranging from −3 to 3 on the latent scale, and person locations (θ_i) were sampled from a normal $(0,2^2)$ distribution to obtain a fair number of extreme simulees. Item discrimination and interthreshold distance were kept constant (with, respectively, $\alpha_g = 1$ and distance between successive τ_{gm} values of 0.4). These values were based on previous studies by Roberts, Donoghue, and Laughlin (2000). For each

participant, the response category with the highest probability was sampled, in that way generating a deterministic data structure.

Figure 1 depicts the OCM diagrams for the nine items in the aforementioned example. In each diagram g, the horizontal axis shows the nine item rank numbers. These rank numbers represent the nine subgroups of respondents who endorsed the corresponding item g. The vertical axis in each diagram g represents the estimated conditional mean responses ($\hat{c}_{gh}$) on item g within the nine subgroups.

In Figure 1 it can be seen that item $v1$ has its peak at the left end of the scale, and the conditional mean response decreases for subgroups choosing items that are more distant from this item. That is, item $v1$ has a monotonically decreasing IRF. In contrast, the IRF of item $v9$ (with its peak at the right end of the scale) is monotonically increasing. The IRF of item $v5$, which is located on the midpoint of the scale, is single-peaked and approximately symmetrical. These IRFs are typical for items that together form an unfolding scale.

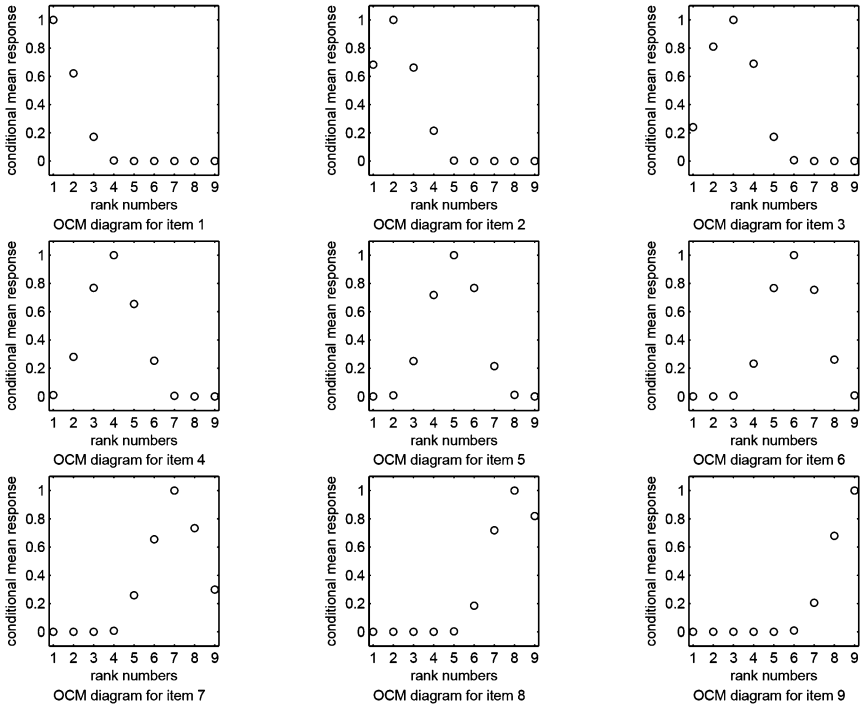

FIGURE 1 Ordered conditional means (OCM) diagrams for simulated responses to nine polytomous items.

Unimodal Smoothing of the OCM Diagrams

We propose a unimodal smoother defined by Eilers (2005) to determine the fit in each diagram. The smoother is based on a nonparametric function and yields the estimated values $\hat{e}_{gh}$ in the OCM diagram of each item g. This approach assumes single-peakedness of the IRFs, unidimensionality, and a hypothesized order of the items on the latent scale. We chose the smoother defined by Eilers because it is the only algorithm (that we know of) that explicitly models a smooth global unimodal shape, whereas other nonparametric curve estimators (cf. Bro & Sidiropoulos, 1998; Gemperline & Cash, 2003) split the unimodal function into monotone left and right parts and do not consider smoothness.

We refer the reader to Eilers (2005) for further technical details and a Matlab routine. We chose to set the roughness penalty, λ (cf. Eilers, 2005, p. 321), to 100 instead of the starting value of 0.1 to enhance smoothness. Given the fact that the current application has relatively few data points compared with the original application of Eilers, a stronger roughness penalty is desired to diminish the influence of individual data points.

Additionally, for this article we included weights for the points in each diagram. That is, because each point h is based on a specific number of observations N_h, we weighted each point h with N_h in the fitting procedure.

Two Measures of Fit for the OCM Diagrams

For each OCM diagram g we determine two measures of fit based on the quantities $\hat{c}_{gh}$, defined in Equation (3), and $\hat{e}_{gh}$, the predicted values resulting from the unimodal smoother.

First, we compute a measure of goodness-of-fit, which is defined for diagram g as the squared correlation (Q_g) between $\hat{c}_{gh}$ and $\hat{e}_{gh}$. Because the predicted values $\hat{e}_{gh}$ resulted from a nonlinear (unimodal) smoothing procedure, this squared correlation cannot be interpreted as the proportion of variance accounted for as is the case in linear regression. Therefore, we avoided the term R-square and used Q_g instead.

Second, we compute the root mean squared error ($RMSE$) as a measure of deviance, which is defined for diagram g as the square root of the (weighted) mean squared error:

$$RMSE_g = \sqrt{\frac{1}{N_+}\sum_{h=1}^{k} N_h(\hat{c}_{gh} - \hat{e}_{gh})^2}, \tag{4}$$

where $N_+ = \sum_{h=1}^{k} N_h$.

To measure the quality of the scale as a whole, we take the average of the values for both measures of fit over all k diagrams. That is, we compute Q, which is

$$Q = \frac{1}{k} \sum_{g=1}^{k} Q_g, \tag{5}$$

and $RMSE$, which is

$$RMSE = \sqrt{\frac{1}{k} \sum_{g=1}^{k} MSE_g}. \tag{6}$$

In general, the higher Q, and the lower $RMSE$, the more the items of the scale show a single-peaked IRF.

Two Measures of Item Fit

As measures of item fit we propose $\Delta Q_{(-h)}$ and $\Delta RMSE_{(-h)}$, which denote the change of, respectively, Q and $RMSE$ when "item h is deleted" from the scale. To compute these statistics, we first compute $Q_{(-h)}$ and $RMSE_{(-h)}$, which are respectively defined by Equations (5) and (6) but with item h discarded from the scale. Subsequently, the resulting new values are compared with the original Q and $RMSE$ values based on k items. Positive values for $\Delta Q_{(-h)}$ and negative values for $\Delta RMSE_{(-h)}$ are indicative of item misfit.

In the following section a diagnostic procedure for examining item fit is explained. Subsequently, cutoff values are derived from a simulation study for both statistics, $\Delta Q_{(-h)}$ and $\Delta RMSE_{(-h)}$, which may be used to decide when to discard an item from the scale.

DIAGNOSTIC PROCEDURE

The following example illustrates the diagnostic procedure and shows how it identifies an item with a deviant IRF. Consider the data from the first example again. However, now we introduce a deviant item to the scale. That is, we made the IRF of item $v3$ a mixture of the IRFs of an item with $\delta = 3$ and $\alpha = 1$ and an item with $\delta = 7$ and $\alpha = 0.25$. In this way the IRF of item $v3$ has a local maximum at the position of item $v7$. Figure 2 shows the OCM diagrams for this data set using the unimodal smoothing procedure to approximate the IRF of each item.

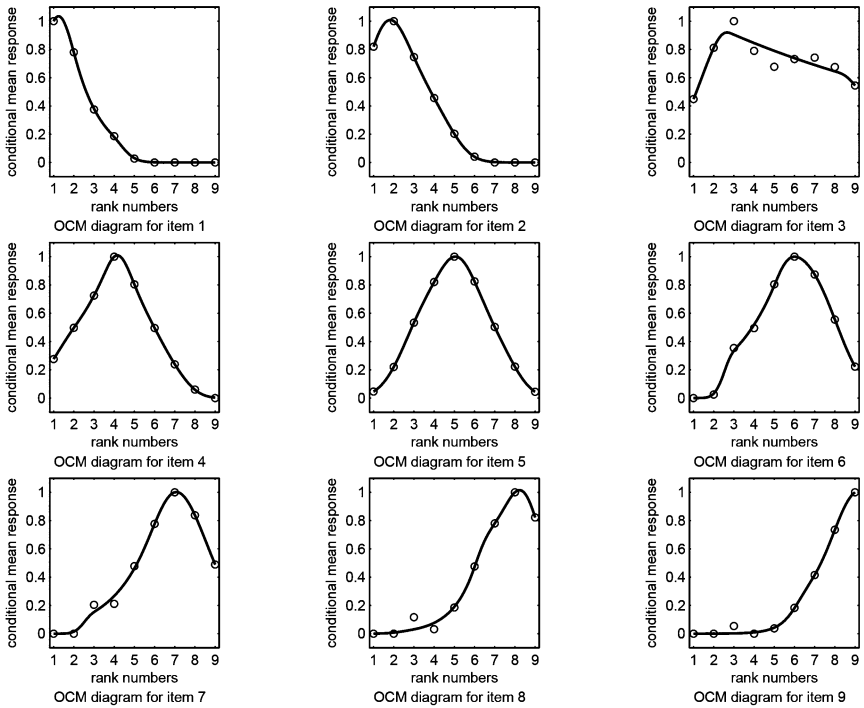

FIGURE 2 Ordered conditional means (OCM) diagrams including a unimodal smoother for simulated responses to 9 polytomous items with item 3 as deviant item.

In Figure 2 it can be seen that the OCM diagram for item $v3$ shows a somewhat deviating pattern. The dots show, as expected, a local maximum around point 7. This means that, when we move to right along the scale, the popularity of item $v3$ does not decrease as much as, for instance, for item $v2$. In practice this would make item $v3$ difficult to scale as it is relatively popular with participants on the left of the midpoint as well with participants on the right of the midpoint.

The scale fit, in terms of Q and $RMSE$, as well as the item fit in terms of both $Q_{(-h)}$ and $RMSE_{(-h)}$ and $\Delta Q_{(-h)}$ and $\Delta RMSE_{(-h)}$ for the aforementioned example are reported in Table 1.

In Table 1 we see that the scale fit can be improved substantially by discarding item $v3$ ($\Delta Q_{(-3)}$ and $\Delta RMSE_{(-3)}$ are, respectively, 0.028 and -0.016). In contrast, discarding any other items had a minimal or negative effect on the scale fit.

TABLE 1
Unimodal Fit for the Nine-Item Scale Displayed in Figure 2

Scale Fit if Item Deleted:	*Q*	*RMSE*	$\Delta Q_{(-h)}$	$\Delta RMSE_{(-h)}$
None	0.972	0.028		
*v*1	0.958	0.030	−0.014	0.002
*v*2	0.967	0.031	−0.005	0.003
*v*3	1.000	0.012	0.028	−0.016
*v*4	0.973	0.027	0.002	−0.001
*v*5	0.982	0.026	0.010	−0.002
*v*6	0.968	0.032	−0.004	0.004
*v*7	0.974	0.029	0.002	0.002
*v*8	0.971	0.030	−0.001	0.002
*v*9	0.965	0.030	−0.007	0.002

Note. $\Delta_{(-h)}$ indicates the change in scale fit when item h is deleted. RMSE = root mean squared error.

SIMULATION STUDY AND REAL EXAMPLES

Analysis of Simulation Data

In this section we present results of a simulation study with a twofold aim; (a) to derive cutoff values for the diagnostic measures $\Delta Q_{(-h)}$ and $\Delta RMSE_{(-h)}$ and (b) to compute the Type I error rate and power for these cutoff values.

The design of the simulation was as follows: First, we used a constant, moderate sample size of $N = 300$. Second, participant distributions were varied to be either $N(0, 2^2)$ or Uniform on (−4.5, 4.5). We chose these parameters to ensure there were not too few respondents with extreme locations relative to the most extreme items. Third, we varied scale length to include either 10 or 15 items. Fourth, all regular items had unimodal IRFs defined by the GGUM (see Appendix B). Item locations (δ_g) were equidistant, ranging from −3 to 3, with constant discrimination ($\alpha_g = 1$) and a constant interthreshold distance of 0.4. Finally, we varied the number of deviant items (0, 1, 2, or 3) and their location along the scale (midpoint, intermediate, or extreme).

We used as deviant either a relatively nondiscriminating IRF or an irregular IRF with an extra local maximum. The discrimination parameter of the nondiscriminating item was set to a much lower value than that of the remaining items. We used the value of $\alpha_g = 0.10$ for the deviant item(s) and $\alpha_g = 1$ for the remaining items in the scale. The "irregular" IRF for a given deviant item was a mixture of an IRF defined by the GGUM for the location δ_g with $\alpha_g = 1$ and a second IRF defined by the GGUM for $\delta_{g\pm3}$ and $\alpha_{g\pm3} = .25$. In that way, the IRF had an extra (local) maximum, which was located at either δ_{g+3}

if the deviant item g has a median rank number or lower, or δ_{g-3} in all other cases.

Results of the Simulation Study

Here we present a brief overview of the main outcomes of the simulation study. Interested readers may contact the corresponding author for a complete report of the simulation study (or see Polak, 2011).

We compared the univariate distributions for both measures of item fit between deviant and regular items. This resulted in identifying $\Delta Q_{(-h)} > 0.025$ and $\Delta RMSE_{(-h)} < -0.005$ as cutoff values for discriminating between items with a deviant IRF and items with a single-peakd IRF.

ANOVAs showed that the performance of neither $\Delta Q_{(-h)}$ nor $\Delta RMSE_{(-h)}$ was affected by participant distribution or scale length. However, the performance of $\Delta Q_{(-h)}$ was moderately affected by type of deviance in the IRF. That is, on the basis of $\Delta Q_{(-h)} > 0.025$ the nondiscriminating items were more often identified than the irregular items.

Furthermore, the performance of $\Delta RMSE_{(-h)}$ was affected by the number of deviant items in the scale. That is, on the basis of $\Delta RMSE_{(-h)} < -0.005$ the deviant items were more often identified when there were relatively few deviant items in the scale.

In Table 2 we report the Type I error and power for the cutoff values $\Delta Q_{(-h)} > 0.025$ and $\Delta RMSE_{(-h)} < -0.005$ in various conditions of the simulation study. The type I error rates were approximated by the average percentage of replications in which $\Delta Q_{(-h)}$ was > 0.025 and $\Delta RMSE_{(-h)}$ was < -0.005 over all $regular$ items. The Type II error rates were approximated by the average percentage of replications in which $\Delta Q_{(-h)}$ was ≤ 0.025 and $\Delta RMSE_{(-h)}$ was ≥ -0.005 over all $deviant$ items. The statistical power of the respective rules of thumb is approximated by 100%–the Type II error rate.

In Table 2 it can be seen that in most conditions the power of our methodology for identifying item misfit is high, whereas in most conditions the Type I error rates remain acceptable. However, for the 10-item scales the Type I error of the $\Delta Q_{(-h)} > 0.025$ cutoff value is $> 5\%$ (but $< 10\%$) when the deviant items are characeterized by a nondiscriminating IRF.

For the $\Delta RMSE_{(-h)} < -0.005$ cutoff value we also see a slightly increased Type I error ($> 5\%$, but $< 10\%$) when the deviant items are characeterized by a nondiscriminating IRF, but only for the conditions with one deviant item. Inspection of the univariate distributions in these conditions indicated that these relatively high Type I error rates were caused by misfit for regular items that were adjacent to the deviant items. However, in these conditions the deviant item still clearly stood out with, on average, values of misfit that were 3 times as

TABLE 2
Type I Error Rates and Power of the Ordered Conditional Means (OCM) Diagnostics $\Delta\ Q_{(-h)}$ and $\Delta\ RMSE_{(-h)}$, and Their Cutoff Values

	10-Item Scale		*15-Item Scale*	
Type/Number of Deviant Items	*Type I Error*	*Power*	*Type I Error*	*Power*
	$\Delta Q_{(-h)} > 0.025$ as threshold for item misfit			
Mixture				
1	3.0%	96.8%	0.1%	67.3%
2	3.8%	97.7%	0.3%	78.3%
3	5.5%	91.9%	0.1%	80.7%
Nondiscriminating				
1	6.2%	99.8%	0.4%	99.8%
2	9.7%	99.7%	0.3%	99.6%
3	5.9%	98.7%	0.4%	98.9%
	$\Delta RMSE_{(-h)} < -0.005$ as threshold for item misfit			
Mixture				
1	5.1%	100%	0.3%	93.0%
2	1.1%	95.6%	0.5%	91.8%
3	1.0%	69.4%	0.1%	85.6%
Nondiscriminating				
1	8.8%	100%	0.2%	91.7%
2	3.6%	99.1%	0.8%	91.2%
3	0.1%	89.1%	0.3%	84.8%

Note. RMSE = root mean squared error.

large as those of the neighboring regular item. We therefore recommend using an iterative procedure in deleting deviant items from the scale, that is, to only delete the item that shows the strongest degree of misfit and to repeat the fitting procedure for the remaining items. This procedure is illustrated in the section Thurstone's Attitude Toward Capital Punishment Scale.

With respect to scale fit, results showed, as expected, a diminishing decline as a function of the number of deviant items in the scale. Values of $Q = 0.970$ and $RMSE = 0.044$ (i.e., the mean scale fit in the conditions including zero deviant items) were found as thresholds of good fit. ANOVA effect sizes showed that the scale fit was affected neither by participant distribution, nor by scale length.

In the following sections we show two applications of the OCM diagrams for evaluating unfolding scales and items. The first example, from the field of personality assessment, focuses on scale evaluation with a given set of items. The second example, from the field of attitude research, focuses on scale construction with the purpose of item selection.

The Developmental Profile: A Bipolar Scale for Personality Development

We analyzed data on personality development collected with the Developmental Profile (DP; Abraham et al., 2001). The DP is an instrument for psychodynamic personality assessment that consists of eight subscales, referred to as developmental levels, each consisting of nine items, referred to as developmental lines.

The items of the DP are scored by a trained professional based on a semistructured interview. A 4-point scale is used to indicate the degree to which each personality characteristic is present (0 = *not present* through 3 = *very clearly present*). The DP of an individual is defined as his total score on each of the eight developmental levels. These total scores are recoded on a scale ranging from 0 through 3.

It is important to note that the items of the DP are organized primarily in cumulative subscales but that results support the notion that underlying these subscales is a substitutive (bipolar) scale ranging from *strongly maladaptive functioning* to *strongly adaptive functioning* (Polak, Van, Overeem-Seldenrijk, Heiser, & Abraham, 2010). In the current article the aim is to investigate the hypothesis that the shape of the response functions for these eight levels (subscales) is single-peaked.

From a developmental perspective, it is presumed, first, that an individual's (total) score pattern shows a peak at that level, which characterizes his or her current level of functioning best. A second presumption is that the individual's scores on the remaining levels will decrease as function of the distance between those levels and his or her "peak" level along the (adaptivity) dimension underlying the DP. As an individual develops, for instance in the course of therapy, his or her peak will shift up the hierarchy of the DP if the persons learns to replace maladaptive behavior with more adaptive behavior.

The current sample consisted of 736 patients who were classified as forensic inpatients ($N = 24$), inpatients ($N = 450$), outpatients ($N = 163$), and normal controls ($N = 99$).

Scale fit and item fit according to the OCM diagnostics are reported in Table 3. Table 3 shows that the overall fit ($Q = 0.785$ and $RMSE = 0.067$) can improve substantially by deleting $v6$. That is, $\Delta Q_{(-6)} = 0.113$ and $\Delta RMSE_{(-6)} = -0.018$ both exceed the cutoff values of, respectively, 0.025 and -0.005.

The OCM diagrams for the eight items representing the various developmental levels of the DP are depicted in Figure 3. In Figure 3 we see that the IRF of item $v6$ shows a local maximum at point $v3$. Furthermore, item $v6$ appears as an outlying point in diagrams 2 and 3.

These results can be interpreted as follows: item $v6$ was relatively most similar to $v3$ and not to items $v5$ and $v7$ as the position of $v3$ in the hierarchy

TABLE 3
Scale Fit and Item Fit According to the Ordered Conditional Means (OCM) Diagnostics for the Eight Levels of the Developmental Profile

		Q	$RMSE$	$\Delta Q_{(-h)}$	$\Delta RMSE_{(-h)}$
	Scale	0.785	0.067		
Scale fit if item deleted:	v1	0.778	0.069	−0.006	0.002
	v2	0.793	0.067	0.008	0.000
	v3	0.834	0.061	0.050	−0.006
	v4	0.829	0.071	0.044	0.003
	v5	0.841	0.071	0.056	0.003
	v6	0.897	0.050	0.113	−0.018
	v7	0.749	0.073	−0.035	0.006
	v8	0.780	0.068	−0.004	0.001

Note. $\Delta_{(-h)}$ indicates the change in scale fit when item h is deleted.

of the developmental profile suggests. Partly, this can be explained from the relation between the constructs that items $v3$ and $v6$ aim to measure, that is, respectively, Egocentricity and Rivalry. Both these types of characteristics are included in the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000) description of the Narcissistic Personality Disorder. However, the DP regards Egocentricity (being unable or unwilling to change one's behavior) as more maladaptive than Rivalry (being preoccupied with proving one's superiority).

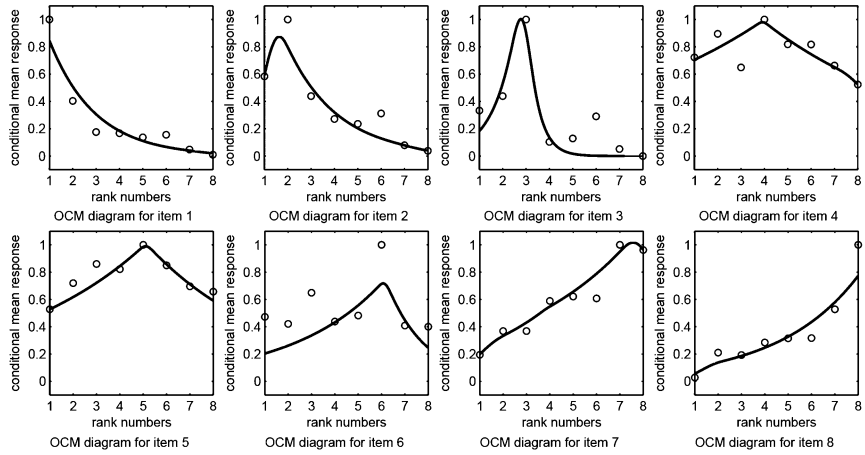

FIGURE 3 Ordered conditional means (OCM) diagrams for the levels of the Developmental Profile.

The results also show that Rivalry, as defined by item $v6$, seems to be a rather general characteristic instead of being apparent only in people with a midpoint position in the DP hierarchy.

The scale fit without item $v6$ ($Q_{(-6)} = 0.897$ and $RMSE_{(-6)} = 0.050$) seems fair given our simulation results for good scale fit (i.e., $Q = 0.970$ and $RMSE = 0.044$). In Figure 3 it can be seen that besides item $v6$ the DP items fit the single-peaked model with shifting peaks fairly well.

Thurstone's Attitude Toward Capital Punishment Scale

In this section we analyze responses to Thurstone's (1932) attitude toward capital punishment scale (see Table 4) obtained by Roberts and Laughlin (1996) from

TABLE 4
Statements of the Capital Punishment Scale (Thurstone, 1932) With Their Original Thurstone Scale Values (*T*)

Statement	*T*
1. Capital punishment is absolutely never justified.	0.0
2. I do not believe in capital punishment under any circumstances.	0.1
3. Capital punishment is the most hideous practice of our time.	0.6
4. Execution of criminals is a disgrace to civilized society.	0.9
5. We can't call ourselves civilized as long as we have capital punishment.	1.5
6. The state cannot teach the sacredness of human life by destroying it.	2.0
7. Capital punishment cannot be regarded as a sane method of dealing with crime.	2.4
8. Capital punishment has never been effective in preventing crime.	2.7
9. Capital punishment is not necessary in modern civilization.	3.0
10. Life imprisonment is more effective than capital punishment.	3.4
11. I don't believe in capital punishment but I'm not sure it isn't necessary.	3.4
12. I think the return of the whipping post would be more effective than capital punishment.	3.9
13. I doesn't matter to me whether we have capital punishment or not.	5.5
14. I do not believe in capital punishment but it is not practically advisable to abolish it.	5.8
15. I think capital punishment is necessary but I wish it were not.	6.2
16. Capital punishment is wrong but is necessary in our imperfect civilization.	6.2
17. Capital punishment may be wrong but it is the best preventative to crime.	7.2
18. Capital punishment is justified only for premeditated murder.	7.9
19. We must have capital punishment for some crimes.	8.5
20. Capital punishment should be used more often than it is.	9.1
21. Capital punishment gives the criminal what he deserves.	9.4
22. Capital punishment is just and necessary.	9.6
23. Any person, man or woman, young or old, who commits murder should pay with his own life.	10.4
24. Every criminal should be executed.	11.0

245 American undergraduates. The scale consists of 24 items (statements), each with a 6-point rating scale (with response categories 0 = *strongly disagree* to 5 = *strongly agree*), varying from strongly against the death penalty to strongly in favor of it.

Roberts and Laughlin (1996) analyzed this data set with GUM (a predecessor of GGUM2004), where they first eliminated items that were unlikely to conform to the unidimensionality assumption based on principal components analysis, leaving 17 items for the analysis. After a first analysis with GUM another 5 items were eliminated based on Wright and Masters' (1982) item-infit t statistic and an item content evaluation. GUM resulted in a final scale consisting of 12 items, where the scale fit was judged as reasonably well based on the squared correlation ($r^2 = .987$) between the average observed and expected values in 70 fitgroups. Besides the item content criterion, no information was presented about the nature of the item misfit of the eliminated items.

In computing the OCM diagnostics we ordered the 24 Thurstone (1932) items according to the original Thurstone scale values (see Table 4). Two pairs of statements (10 and 11, 15 and 16) had equal Thurstone scale values but unequal GUM scale values δ_h (as reported in Roberts & Laughlin, 1996); those pairs were ordered according to the GUM estimates (δ_h).

Table 5 gives the OCM fit measures for the 24 Thurstone items. The initial scale fit based on all items was $Q = 0.710$ and $RMSE = 0.084$. After five iterations the item statistics were all below the cutoff values. Items 12, 13, 14, and 24 were identified as deviant items. The final scale fit based on 20 items was $Q = 0.832$ and $RMSE = 0.079$, which indicates a reasonable fit given the outcomes of the simulation study. Items 12, 13, 24, and 14 were subsequently identified based on $\Delta Q_{(-h)}$, whereas the corresponding values of $\Delta RMSE_{(-h)}$ did not exceed the cutoff value of -0.005 (although $\Delta RMSE_{(-h)}$ was negative in all cases).

Figure 4 shows the OCM diagrams for all 24 statements. In Figure 4 it can be seen that the peaks of the estimated IRFs move from left to right, with, for instance, the IRF of item 15 ("I think capital punishment is necessary, but I wish it were not") as a typical example of an IRF that first increases and than decreases again. We see that most OCM diagrams show the assumed single-peaked pattern and apparently discriminate well between different attitudes.

Moreover, Figure 4 shows that the discarded items indeed have a deviant IRF. The IRFs of items 12 and 13 are slightly irregular but even more so nondiscriminating. For instance, the item wording of item 12, "I think the return of the whipping post would be more effective than capital punishment," was originally intended to reflect a moderate position on the attitude continuum. However, in the current sample, agreement with this statement seems independent of one's position on the underlying attitude continuum.

TABLE 5
Scale Fit and Item Fit According to the Ordered Conditional Means (OCM) Diagnostics for the 24 Capital Punishment Items

		Iteration 1		*Iteration 2*		*Iteration 3*		*Iteration 4*		*Iteration 5*	
		Q	*RMSE*	*Q*	*RMSE*	*Q*	*RMSE*	*Q*	*RMSE*	*Q*	*RMSE*
		.710	*.084*	*.742*	*.082*	*.775*	*.082*	*.807*	*.080*	*.832*	*.079*
Item	δ_h	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$	$\Delta_{(-h)}$
1	−1.43	−.014	.002	−.014	.002	−.013	.002	−.011	.002	−.010	.002
2	−1.86	−.010	.000	−.001	.000	−.009	.000	−.008	−.000	−.008	−.000
3	.	.006	−.002	.009	−.003	.010	−.003	.011	−.003	.013	−.004
4	−1.18	−.011	.001	−.009	.001	−.007	.001	−.006	.001	−.005	.001
5	−1.30	−.015	.001	.014	.001	−.012	.001	−.010	.001	−.010	.001
6	.	−.021	.003	−.020	.003	−.018	.003	−.011	.003	−.019	.003
7	.	−.015	.000	−.014	.001	−.013	.000	−.010	.000	−.010	.000
8	.	.002	.000	.004	−.000	.006	−.001	.007	−.001	.005	−.001
9	−1.38	−.003	.000	−.002	.001	−.000	.000	.001	−.000	.001	−.000
10	−1.08	−.007	.001	−.006	.001	−.004	.001	−.002	.000	.001	.000
11	−.83	.006	.000	.007	.000	.008	.000	.010	.000	.015	−.000
12	.	**.032**	−.001	.	.	.	.	.	.	.	.
13	.	.030	−.000	**.033**	−.000	.	.	.	.	.	.
14	−.60	.024	−.000	.024	−.000	.023	−.000	**.026**	−.001	.	.
15	.	.007	.005	.008	.005	.013	.004	.016	.004	.006	.005
16	.	.003	.001	.005	.001	.006	.001	.008	.001	.010	.001
17	.	.002	.001	.003	.001	.005	.001	.006	.001	.008	.001
18	.	.010	.000	.011	.000	.012	.000	.014	.000	.016	.000
19	2.50	−.003	.008	−.002	.007	−.001	.007	.001	.007	.004	.007
20	2.70	−.015	.004	−.014	.004	−.013	.004	−.010	.004	−.008	.004
21	2.46	−.011	.004	−.011	.004	−.009	.004	−.008	.004	−.007	.004
22	2.45	−.011	.005	−.009	.005	−.009	.005	−.011	.005	−.006	.005
23	.	.003	.001	.005	.000	.007	.000	.008	.000	.007	.000
24	.	.026	−.001	.029	−.001	**.032**	−.001	.	.	.	.

Note. Bold and underlined values exceed the proposed cut-off criteria; the corresponding item is discarded in the next iteration; δ_h = the graded unfolding model (GUM) item location estimate from Roberts and Laughlin (1996); $\Delta_{(-h)}$ indicates the change in scale fit when item h is deleted; RMSE indicates root mean squared error.

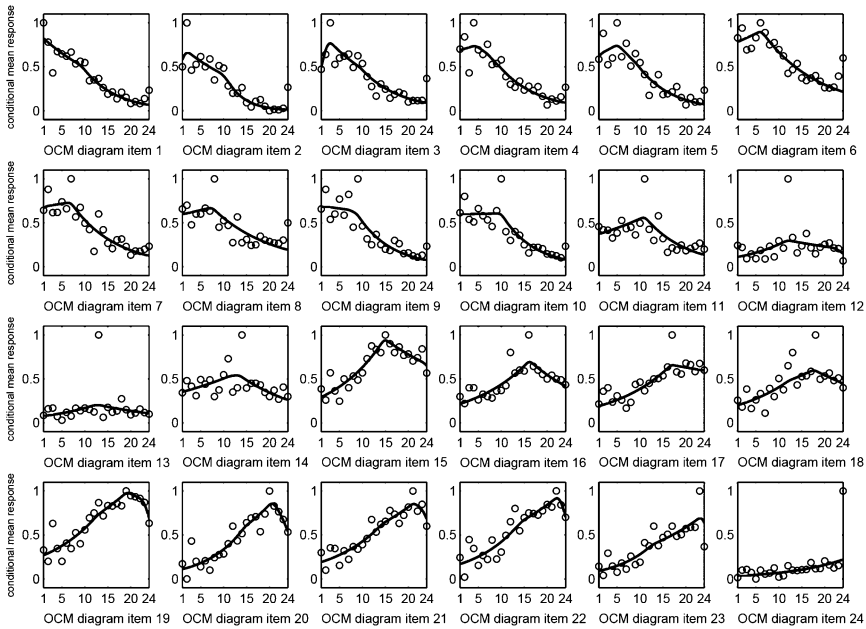

FIGURE 4 Ordered conditional means (OCM) diagrams for the 24 items of Thurstone's attitude toward capital punishment scale.

The IRF of item 14 is nondiscriminating but also shows an extra peak around item 9, which indicates ambiguity. Item 24 is again strongly nondiscriminating as it is unpopular over the entire range of the scale. Its wording is "Every criminal should be executed," which indeed seems far too strong, especially given this (student) sample.

In conclusion, this example showed that the OCM methodology is an informative and useful method for scale construction. Furthermore, this methodology provided insight to the nature of item misfit and resulted in discarding fewer items than in Roberts and Laughlin's (1996) approach. It should be noted that the GUM is a model that explicitly estimates item and person locations and is in that sense incomparable to the OCM methodology (which only aims at evaluating the IRFs of a set of items with a specified order).

However, it turned out that in Roberts and Laughlin's (1996) approach 7 items (6, 7, 8, 15, 16, 17, 18, 23) were discarded that in the current analysis did not show substantial misfit. The items that were identified as deviant by the OCM methodology were also among the discarded items in Roberts and Laughlin's approach with the only exception item 14, which seemed relatively

nondiscriminating and irregular (see Figure 4, diagram 14) but was included (near the midpoint) in the GUM scale.

DISCUSSION

In this article we proposed a model-free diagnostic for single-peakedness of item responses. We introduced the ordered conditional means (OCM) methodology, which is a generalization of Thurstone & Chave's (1929) criterion of irrelevance that was originally developed for dichotomous attitude items. One advantage of the OCM methodology is that it allows for the inclusion of ordered polytomous (graded) responses. Additionally, to the graphical OCM approach we included a unimodal smoother, which we adopted from the field of chemical modeling (Eilers, 2005). Combining the OCM method with the smoothing approach allows for the examination of IRFs in a model-free fashion.

Furthermore, we defined measures of fit for the application of the smoothing procedure to the OCM diagrams. This yielded the indices Q and $RMSE$ as measures of aggregate scale fit. Item fit for a given item h was determined by the change in scale fit "if item h deleted," that is, $\Delta Q_{(-h)}$ and $\Delta RMSE_{(-h)}$. Based on a simulation procedure $\Delta Q_{(-h)} \geq 0.025$ and $\Delta RMSE_{(-h)} \leq -0.005$ were proposed as cutoff values for identifying deviant items. These cutoff values for item misfit showed high power rates while at the same time the Type I error rates remained acceptable in most conditions. Values of $Q = 0.970$ and $RMSE = 0.044$ were found as thresholds of good scale fit.

In the simulation procedure the aim was to resemble realistic empirical conditions. This included using as deviant items two moderate violations of single-peaked IRFs. That is, we used as deviant either a relatively flat (i.e., nondiscriminating) IRF or an irregular IRF with one extra local maximum. We argue that, as the OCM diagnostics were shown to be sensitive to these moderate violations of single-peakedness, it is likely that the diagnostics will also detect more extreme cases of misfit. An example of a more extreme case of misfit is an item with a single-dipped IRF. Single-dipped IRFs were used by Post (1992) in a simulation procedure similar to one performed in the current study. However, Post was criticized by Sijtsma (1995), who called this type of deviance "highly unrealistic" and recommended using "irregularly shaped" instead.

The OCM method has several advantages compared with unfolding IRT-based item fit statistics. First, as it is not based on a parametric model, it allows the approximation of the IRFs with less stringent assumptions. Where item misfit may cause nonconvergence for unfolding IRT methods, the currently proposed methodology always resulted in a solution including all items, thus providing a practical researcher with useful information for deciding to discard or improve specific items. Second, as the proposed method is computationally

relatively simple, it does not require large samples (we showed that the OCM method works well with samples of $N = 300$). Third, we provided a graphical approximation of the IRFs together with measures of deviance from single-peakedness. As we also derived cutoff scores for these measures of deviance, the approach is relatively straightforward for practical users.

A possible weakness of the current methodology is that it depends on a sound hypothesis concerning the true item ordering. Failure in specifying the correct item ordering will also result in misfit. The OCM methodology could be combined with subject matter expert judgments concerning the item order. This classical Thurstonian method circumvents the estimation problems (1928). In this article, we suggested some data analytical methods for ordering the data besides unfolding IRT. In particular, we recommend performing CA on the raw data to derive an optimal item ordering (see Polak et al., 2009). However, regardless of the model used, there may be conditions in which it is difficult to estimate the item ordering, for instance, when items vary in how strongly they discriminate among respondents. Particularly, when several items appeal to respondents from a broader range of locations on the continuum, the (observed) preferred item ordering will vary across respondents.

In the current simulation procedure we chose participant locations that were relatively spread out compared with the item locations. In this way, we ensured that there were not too few respondents with extreme locations relative to the most extreme items. As the OCM methodology selects subsets of respondents on the basis of maximum agreement, it might be difficult to find subsets endorsing items with extreme locations. Therefore, one could consider selecting a larger subset by allowing a user-defined cutoff value for agreement, for instance, $Z_{ih} \geq M - 1$ (cf. Van Schuur 1992, p. 65).

The examples from the fields of personality assessment and attitude research showed that the OCM methodology is useful for both scale evaluation and item selection. It was demonstrated that the OCM diagrams provide useful information about the nature of item misfit. However, future research must establish the generalizability of the cutoff scores (under different response models, other types of item misfit, and varying levels of sample size). Hence, the proposed cutoff values must not be used rigidly.

But again, we stress that the proposed OCM approach and the existing IRT approaches might be used at different stages of the process of scale construction. As we showed in the example with the Thurstone attitude scale (1932), the current preselection steps that are required to obtain GGUM estimates might be too rigorous. Future research could show whether the OCM method might result in a more optimal preselection of items that fit the GGUM. Furthermore, it is currently investigated under which conditions the combination of using CA to order the items together with the OCM method to evaluate item fit would be a useful alternative to unfolding IRT methods such as GGUM.

We used a unimodal smoother defined by Eilers (2005) to determine the fit in each diagram. To obtain a more parsimonious model as an estimate for the IRF, the OCM diagrams could also be modeled with a parametric unimodal function, such as the Gaussian. An advantage of such a procedure is that it has only one parameter for curve width, which could be interpreted as a measure for item discrimination. A disadvantage is that this is a more restrictive model, which imposes constraints such as symmetry of the IRF, that might not be realistic for real data. When trying out the Gaussian in the simulation study that was presented in this article, we found the same pattern of results as with the unimodal smoother, although with overall poorer and more variable values for both measures of fit.

Although this article focuses on item fit, the OCM methodology could be generalized to measure person fit. For that purpose, instead of the estimated conditional mean responses, the actual responses of every person could be plotted against the item ranks. Note that for a data matrix with ordered (single-peaked) items the expected score pattern in the rows is single-peaked as well. Several authors (e.g., Emons, Sijtsma, & Meijer, 2004; Sijtsma & Meijer, 2001) showed that the person response function is an important tool in person-fit research. Analogous to the interpretation of the conditional mean responses and the OCM diagrams proposed in this article, we could interpret the score pattern (standardized with respect to the maximum score) within each row of the data matrix as a rough estimate of the person response function. Accordingly, we could assess the person fit by fitting the smoother as proposed in this article.

Altogether, we think the OCM methodology provides useful diagnostics for both item fit and overall scale fit for single-peaked response items.

REFERENCES

Abraham, R. E., Van, H. L., Van Foeken, I., Ingenhoven, T. J. M., Tremonti, W., de Vries, I. P., ... Spinhoven, P. (2001). The Developmental Profile, *Journal of Personality Disorders, 15,* 457–473.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49,* 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17,* 253–276.

Andrich, D., & Styles, I. M. (1998). The structural relationship between attitude and behavior statements from the unfolding perspective. *Psychological Methods, 3,* 454–469.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Ashby, F. G., & Ennis, D. M. (2002). A Thurstone-Coombs model of concurrent ratings with sensory and liking dimensions. *Journal of Sensory Studies, 17,* 43–59.

Benzécri, J.-P. (1973). *L' analyse des données 1. La Taxinomie, 2. L'analyse des correspondances* [Data analysis 1. The taxonomy, 2. Correspondence analysis]. Paris, France: Dunod.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Bro, R., & Sidiropoulos, N. D. (1998). Least squares algorithms under unimodality and non-negativity constraints. *Journal of Chemometrics, 12,* 223–247.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49,* 743–748.

Carter, N. T., & Zickar, M. J. (2011). The influence of dimensionality on parameter estimation accuracy in the generalized graded unfolding model. *Educational and Psychological Measurement, 71,* 765–788.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal-point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88–106.

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12,* 83–97.

Dalal, D. K., Withrow, S., Gibby, R. E., & Zickar, M. J. (2010). Six questions that practitioners (might) have about ideal-point response process items. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3,* 498–501.

De Leeuw, J., & Mair, P. (2009). Simple and canonical correspondence analysis using the R package anacor. *Journal of Statistical Software, 31,* 1–21.

DeMars, C. E. (2004). Type I error rates for generalized graded unfolding model fit indices. *Applied Psychological Measurement, 28,* 48–71.

Eilers, P. H. C. (2005). Unimodal smoothing. *Journal of Chemometrics, 19,* 317–328.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39,* 1–35.

Gemperline, P. J., & Cash, E. (2003). Advantages of soft versus hard constraints in self-modeling curve resolution problems: Alternating least squares with penalty functions. *Analytical Chemistry, 75,* 4236–4243.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester, UK: Wiley.

Greenacre, M. J. (1993). *Correspondence analysis in practice.* London, UK: Academic Press.

Habing, B., Finch, H., & Roberts, J. S. (2005). A Q3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement, 6,* 457–471.

Heiser, W. J. (1981). *Unfolding analysis of proximity data* (Unpublished doctoral dissertation). Leiden University, The Netherlands.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied psychological measurement, 15,* 153–169.

Hubert, L., Arabie, P., & Meulman, J. (1998). The representation of symmetric proximity data: Dimensions and classifications. *The Computer Journal, 41,* 566–577.

Ihm, P. (2005). A contribution to the history of seriation in archaeology. In C. Weihs & W. Gaul (Eds.), *Classification: The ubiquitous challenge. Proceedings of the 28th annual Conference of the Gesellschaft fur Klassifikation* [Classification society], *University of Dortmund, March 9-11, 2004* (pp. 307–316). Berlin, Germany: Springer-Verlag.

Javaras, K. N., & Ripley, B. D. (2007). An unfolding latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102,* 454–463.

Johnson, M. S. (2006). Nonparametric estimation of item and respondent locations from unfolding-type items. *Psychometrika, 71,* 257–279.

Maydeu-Olivares, A., Hernandez, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41,* 445–471.

Meulman, J. J., & Heiser, W. J. (2004). *SPSS Categories 13.0.* Chicago, IL: SPSS Inc.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'Hara, R. B., Simpson, G. L., ... Wagner, H. (2010). Vegan: Community Ecology Package [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=vegan (R Package Version 1.17-3).

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Polak, M. G. (2011). *Item analysis of single-peaked response data: The psychometric evaluation of bipolar measurement scales* (Unpublished doctoral dissertation). Leiden University, The Netherlands.

Polak, M. G., Heiser, W. J., & De Rooij, M. (2009). Two types of single-peaked data: Correspondence analysis as an alternative to principal component analysis. *Computational Statistics and Data Analysis, 53,* 3117–3128.

Polak, M. G., Van, H. L., Overeem-Seldenrijk, J., Heiser, W. J., & Abraham, R. E. (2010). The Developmental Profile: Validation of a theory driven instrument for personality assessment. *Psychotherapy Research, 20,* 259–272.

Post, W. J. (1992). *Nonparametric unfolding models: A latent structure approach.* Leiden, The Netherlands: DSWO Press.

Roberts, J. S. (2008). Modified Likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement, 32,* 407–423.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26,* 192–207.

Roberts, J. S., Fang, H., Ciu, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 30,* 64–65.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231–255.

Roberts, J. S., & Thompson, V. M. (2011) Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35,* 259–279.

SAS Institute Inc. (2008). *SAS/STAT 9.2 user's guide.* Cary, NC: Author.

Sijtsma, K. (1995). Review of the book *Nonparametric unfolding models: A latent structure approach,* by W. J. Post. *Journal of Classification, 12,* 153–156.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191–208.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model-data fit of ideal point and dominance models. *Applied Psychological Measurement, 35,* 280–295.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94,* 1287–1304.

Ter Braak, C. J. F., & Prentice, I. C. (1988). A theory of gradient analysis. *Advances in Ecological Research, 18,* 271–317.

Ter Braak, C. J. F., & Prentice, I. C. (2004). A theory of gradient analysis. *Advances in Ecological Research, 34,* 235–282.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34,* 273–286.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.
Thurstone, L. L. (1932). Motion pictures and attitudes of children. Chicago, IL: University of Chicago Press.
Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church.* Chicago, IL: University of Chicago Press.
Van Schuur, W. H. (1992). Nonparametric unidimensional unfolding for multicategory data. *Political Analysis, 4,* 41–74.
Van Schuur, W. H., & Kiers, H. A. L. (1949). Why factor analysis often is the wrong model for analysing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18,* 97–110.
Van Schuur, W. H., & Post, W. J. (1998). *MUDFOLD users manual.* Groningen, The Netherlands: iec. ProGamma.
Weekers, A. M. (2009). *Modeling typical performance measures* (Unpublished doctoral dissertation). University of Twente, Enschede, The Netherlands.
Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items. *European Journal of Psychological Assessment, 24,* 65–77.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.
Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

APPENDIX A

Correspondence Analysis as Approach to Scaling Single-Peaked Items

In the following we give the computational details and the rationale of correspondence analysis (CA) of item response data. The formulation is adapted from Greenacre (1993).

Let $\mathbf{Z}$ denote the original data matrix, where the entries z_{ij} indicate the observed response of participant i ($i = 1, \ldots, n$) to item j ($j = 1, \ldots, k$). The responses are considered measures of association strength between the row entry (here: participant i) and column entry (here: item j). The association measure is assumed to be some nonnegative quantity, where lack of association (for instance, a *strongly disagree* response to an attitude item with a graded response scale) is indicated by a zero entry.

It is algebraically simpler to work with the so-called correspondence matrix $\mathbf{P}$, with elements $p_{ij} = z_{ij}/z_{++}$, where the index $+$ indicates the sum over the omitted index. From $\mathbf{P}$ we compute the matrix $\mathbf{D}$, with standardized deviations from independence, d_{ij}, where

$$d_{ij} = (p_{ij} - p_{i+}p_{+j})/\sqrt{p_{i+}p_{+j}}.$$

Note that if the participants and items are independent (which means that the participants' ratings of the various items cannot be explained from their mutual distances on one or only a few latent scales(s)), an element p_{ij} equals the product $p_{i+}p_{+j}$. By weighing the deviations from independence with the respective marginals p_{i+} and p_{+j}, we obtain the matrix $\mathbf{D}$ of standardized deviations from independence.

For $\mathbf{D}$ we compute the singular value decomposition: $\mathbf{D} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T$, where $\mathbf{U}$ is the matrix of left singular vectors, with elements u_{is}, $s = 1, \ldots, q$, with $q = \min(n-1, k-1)$; $\mathbf{\Delta}$ is a diagonal matrix with positive singular values l_s in descending order along the diagonal; and $\mathbf{V}$ is the matrix with right singular vectors, with elements v_{js}.

The aim of CA is to find a lower dimensional representation of $\mathbf{D}$. The CA estimated participant location $\hat{\theta}_{is}$ and estimated item location $\hat{\delta}_{js}$ on dimension s can be expressed as, respectively,

$$\hat{\theta}_{is} = l_s^{1-a} \cdot u_{is} / \sqrt{p_{i+}}$$

and

$$\hat{\delta}_{js} = l_s^{a} \cdot v_{js} / \sqrt{p_{+j}}.$$

There are three choices for a in common usage, namely, $a = 0$, 1, or 1/2 (also referred to as, respectively, row principal, column principal, and symmetrical normalization). With $a = 0$ the participant locations $\hat{\theta}_i$ are weighted averages of the sample locations $\hat{\delta}_j$ (which is called by Benzécri, 1973, "le principe barycentrique" [the Barycentric principle]), which is the choice of normalization that corresponds to the notion of the participant scaling in Thurstone's (1928) method.

For one-dimensional data, only the first left and right singular vectors and the first singular value are used to determine, respectively, the participant and item location estimates. The quality of the lower dimensional representation of the data is derived from the singular values l_s and is expressed as the percentage of the total inertia that is explained by each dimension. The total inertia of the data table is the chi-square statistic divided by n, which can be written as

$$\chi^2/n = \sum_{i=1}^{n} p_{i+} \sum_{j=1}^{k} (z_{ij}/z_{i+} - p_{+j})^2 / p_{+j}.$$

The total inertia of the data table can be regarded as the weighted average of the squared deviations between the participants' profiles (the participants' scores proportional to their total score) and the average score profile. Hence, it can be

thought of as the amount of variation among the participants' score patterns. The total inertia of the data table is identical to

$$\sum_{s=1}^{q} l_s^2,$$

where l_s^2 (which equals the eigenvalue λ_s) is referred to as the principal inertia of dimension s. The percentage of inertia explained by dimension s is

$$100 \times l_s^2 / \sum_{s=1}^{q} l_s^2.$$

The contribution of item point $\hat{\delta}_{js}$ to the inertia of dimension s is

$$p_{+j} \hat{\delta}_{js}^2 / l_s^2,$$

or, equivalently, of participant point $\hat{\theta}_{is}$, the contribution to the inertia of dimension s is

$$p_{i+} \hat{\theta}_{is}^2 / l_s^2.$$

APPENDIX B

GGUM: An Unfolding IRT Model for Scaling Single-Peaked Items

The generalized graded unfolding model (GGUM) is a parametric item response model that incorporates features such as variable item discrimination and variable threshold parameters for the response categories. The model assumptions are as follows: existence of a latent trait (i.e., unidimensionality); local independence; and symmetric (around the item location), bell-shaped item response functions. The GGUM allows for binary or graded responses. One premise of the GGUM is that for each person there are two *subjective responses* associated with each observable response. These subjective responses can be seen as two distinct reasons for a person's response. For instance, when a person strongly disagrees with a certain item this could be for either of two reasons. If on the underlying dimension the item is located more to the right extreme than the person, the person disagrees *from below* the item. However, if the item is located more to the left extreme than the person, the person disagrees *from above* the item. The probability that a person will respond using a particular observable answer

category is defined as the sum of the probabilities associated with the two corresponding subjective responses. Specifically, the model has the form

$$P(Z_{ig} = z|\theta_i) = \frac{\exp\{\alpha_g[z(\theta_i - \delta_g) - \sum_{m=0}^{z} \tau_{gm}]\} + \exp\{\alpha_g[(S - z)(\theta_i - \delta_g) - \sum_{m=0}^{z} \tau_{gm}]\}}{\sum_{\omega=0}^{M} \left(\exp\{\alpha_g[\omega(\theta_i - \delta_g) - \sum_{m=0}^{\omega} \tau_{gm}]\} + \exp\{\alpha_g[(S - \omega)(\theta_i - \delta_g) - \sum_{m=0}^{\omega} \tau_{gm}]\} \right)},$$

where Z_{ig} is the observed response of participant i $(i = 1, \ldots, n)$ to item g $(g = 1, \ldots, k)$ with $Z_{ig} = z$, where $z = 0, 1, \ldots, M$ with $z = 0$ indicating the strongest level of disagreement, $z = M$ indicating the strongest level of agreement, $S = 2M + 1$, θ_i is the location of person i, δ_g is the location of item g (on the same metric as θ), α_g is the discrimination of item g, and τ_{gm} is the relative location of response category m within item g.

The GGUM model parameters can be estimated with GGUM2004 (Roberts et al., 2006), which works as follows: item parameters are estimated using a marginal maximum likelihood approach (Bock & Aitken, 1981; Bock & Lieberman, 1970). The algorithm is based on an expectation maximization strategy, which is used to solve the likelihood equations for the item parameters, δ_g, α_g, and τ_{gm}. Participant parameter estimates are obtained by using an expected a posteriori procedure. For a more extensive treatment of the GGUM2004 estimation procedure, see Roberts et al. (2002). Free copies of the program are readily available to readers at http://www.psychology.gatech.edu/unfolding/FreeSoftware.html