Advance Access publication September 12, 2015

Political Analysis (2015) 23:550–563
doi:10.1093/pan/mpv020

Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest

Daniel L. Oberski
Department of Methodology and Statistics, Tilburg University, Room P 1107, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: d.oberski@tilburguniversity.edu (corresponding author)

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University, Room P 1114, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: j.k.vermunt@tilburguniversity.edu

Guy B. D. Moors
Department of Methodology and Statistics, Tilburg University, Room P 1110, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: g.b.d.moors@tilburguniversity.edu

Edited by R. Michael Alvarez

Many variables crucial to the social sciences are not directly observed but instead are latent and measured indirectly. When an external variable of interest affects this measurement, estimates of its relationship with the latent variable will then be biased. Such violations of ''measurement invariance'' may, for example, confound true differences across countries in postmaterialism with measurement differences. To deal with this problem, researchers commonly aim at ''partial measurement invariance''; that is, to account for those differences that may be present and important. To evaluate this importance directly through sensitivity analysis, the ''EPC-interest'' was recently introduced for continuous data. However, latent variable models in the social sciences often use categorical data. The current paper therefore extends the EPC-interest to latent variable models for categorical data and demonstrates its use in example analyses of U.S. Senate votes as well as respondent rankings of postmaterialism values in the World Values Study.

1 Introduction

Latent variable models for categorical data are commonly used in the social and behavioral sciences, and include well-known special cases such as latent class analysis, item response theory (IRT), and ordinal factor analysis. Examples of such analyses include ideological positions of U.S. Senators measured by Yea/Nay votes (Poole and Rosenthal 1985), public tolerance for nonconformity measured by categorical questions in a survey (McCutcheon 1985), and postmaterialism measured by respondents' rankings of values (Inglehart 1981; Moors and Vermunt 2007).

Primary scientific interest in latent variable models often focuses on the relationship such latent variables have with external variables; for example, on the polarization of ideology by political party, on education and cohort differences in tolerance, or on cross-country differences in postmaterialism. However, if the measurement of the latent variable differs over values of such external variables, relationship estimates of interest may be biased (Steenkamp and Baumgartner 1998). Thus, ''measurement invariance'' is needed to estimate such relationships.

Authors' note: The authors are grateful to two anonymous reviewers, the editor, Katrijn van Deun, Lianne Ippel, Eldad Davidov, Zsuzsa Bakk, Verena Schmittmann, and participants of EAM2014 for their comments. Replication materials for this article can be obtained from http://dx.doi.org/10.7910/DVN/I7Y3G2, Harvard Dataverse, V1. Supplementary materials for this article are available on the *Political Analysis* Web site.

© The Author 2015. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

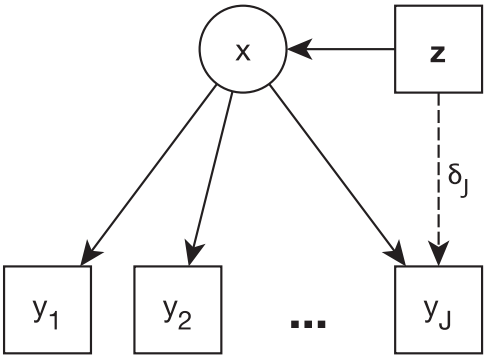


Fig. 1 Graph of a latent variable model. The latent variable x is shown in a circle to indicate that it is unobserved.

For this reason, whenever the aim is to compare values of the latent variable, it is common practice to perform "measurement invariance testing" (see Vandenberg and Lance 2000; Schmitt and Kuljanin 2008 for reviews), a practice also known as testing for "differential item functioning" (DIF) in the IRT literature (Holland and Wainer 1993); that is, detecting any observed indicators that might violate measurement invariance and accounting for these violations in the model. The aim of measurement invariance testing is to ensure the comparability of latent variable scores over values of an external variable. However, currently existing procedures do not necessarily reach this aim because they do not account for the substantive impact of violations of measurement invariance on the relationship parameters of interest (Oberski 2014).

To this end, Oberski (2014) suggested supplementing measurement invariance testing for linear structural equation models (SEMs) with sensitivity analysis. Sensitivity analysis investigates directly the impact of measurement invariance violations on the relationship parameters of interest. It can be performed by fitting many alternative models, with the disadvantage that this process will sometimes be infeasible. Oberski (2014) therefore also introduced the EPC-interest, a measure that approximates the change in parameters of interest without the need to fit alternative models. Since EPC-interest was only introduced for continuous data linear SEMs, however, it is not applicable to categorical data analyses such as those described above.

In this article, we therefore extend the EPC-interest approach to measurement invariance testing to the case of categorical data latent variable models. The extended measure is evaluated in a small simulation study, and its use is demonstrated in two example applications. In the first of these applications, the polarization of the 90th U.S. Senate is examined by applying an ideal point model; the second application analyzes the rankings of value priorities given by 67,568 WVS respondents in forty-eight countries.

Section 2 explains categorical data latent variable models and the problem of measurement invariance. The EPC-interest is introduced in Section 2.2 as an approximation to the sensitivity analysis approach to measurement invariance testing, and a small simulation study evaluates the approximation in Section 2.3. Sections 3 and 4 then elaborate the two example applications, while Section 5 concludes. Program inputs and data for the simulation and examples discussed in this article are provided in the online appendix (Oberski 2015).[1]

2 Measurement Invariance in Categorical Data

2.1 *The Problem of Measurement Invariance*

Many important variables in the social sciences are not or cannot be directly observed, but are instead latent (Bollen 2002) and measured using a vector of multiple indicators. Often this "measurement model" is not, however, of primary interest to the researcher, but rather the relationship between the latent variable and some covariate vector is: the "structural model."

[1] http://dx.doi.org/10.7910/DVN/I7Y3G2.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Figure 1 shows this relationship as an arrow between the observed covariate $\mathbf{z}$ and latent variable x, measured by observed variables $\mathbf{y}$. For example, $\mathbf{z}$ could contain a set of dummy variables indicating a respondent's country or gender, so that the relationship simply compares values of x across countries or genders. Alternatively, interest could focus on the influence of continuous covariates such as GDP or age on the latent x. Of course, the problem is that x is not observed, but only $\mathbf{y}$ and $\mathbf{z}$ are.

Latent variable models generally attack this problem by assuming that, while the latent variable may depend on the covariate, its measurement should not; that is, the indicators should be "measurement invariant":

$$p(y_j|x, \mathbf{z}) = p(y_j|x). \tag{1}$$

When Equation (1) holds for all J indicators, this is called "full measurement invariance" (Meredith 1993). A violation of measurement invariance is sometimes termed "DIF," since the "items" y_j have differing conditional probability distributions given x for differing values of $\mathbf{z}$ (Mellenbergh 1989). Intuitively, measurement invariance is necessary to identify the structural model because if all indicators are caused by both x and $\mathbf{z}$, then there is no way of knowing whether an observed indicator difference over values of the covariate is due to a covariate effect on x or to a difference over the covariate in the way x is measured. For example, when comparing countries, an observed cross-country difference in anti-immigrant attitudes might be a substantive difference or it might equally well be explained as differing answer tendencies over countries.

This identification problem only occurs, however, when there are no measurement-invariant indicators at all. A single invariant indicator can disentangle measurement from substantive differences in the other, non-invariant, indicators. Full measurement invariance is therefore not necessary to identify the structural model, but "partial measurement invariance" is (Byrne, Shavelson, and Muthén 1989). The standard practice is to search for non-invariant indicators and either remove them or parameterize their differential functioning (Holland and Wainer 1993). A common way of doing so when the observed variables are categorical with K categories is a logistic regression,

$$P(Y_j = k) = \frac{\exp(\tau_{jk} + \lambda_{jk}x + \delta_{jk}\mathbf{z})}{\sum_{m\in\{1..K\}} \exp(\tau_{jm} + \lambda_{jm}x + \delta_{jm}\mathbf{z})} \tag{2}$$

(Mellenbergh 1989; Kankaraš, Moors, and Vermunt 2010). It can be seen in Equation (2) that setting $\delta_{jk} = 0$ corresponds to measurement invariance.[2] This means that it is possible to estimate a model in which $\delta_{jk} \neq 0$ for *some* indicators, while δ_{jk} is kept at zero, that is, partially measurement invariant, for others. Figure 1 shows this possible direct effect from covariate to indicator as a dashed arrow.

Some form of measurement invariance is needed to estimate the structural model of interest. Deciding which form, that is, selecting a model, is therefore crucial. It is common to test for full or partial measurement invariance using various fit measures available for this purpose (Byrne, Shavelson, and Muthén 1989; Hu and Bentler 1998; Cheung and Rensvold 2002; Chen 2007; Saris, Satorra, and Van der Veld 2009). However, this approach has recently been shown to have an unfortunate disadvantage: when a violation is detected, it need not have seriously affected the parameter of interest, and when an invariance model is selected, substantial bias in the parameter of interest may still remain (Oberski 2014). In short, while measurement invariance is an important assumption in latent variable modeling, the existing practice does not directly account for the effect that violations of this assumption have on the parameters of interest.

To complement measurement invariance testing, the EPC-interest was therefore recently introduced by Oberski (2014) for continuous data SEMs. The EPC-interest assesses what would happen if a particular possible direct effect were freed. Rather than assessing the size of this direct

[2] An interaction term $\delta^*_{jk}\mathbf{z}x$ could be added to Equation (2) but has been omitted for clarity.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

effect itself, however, it assesses its impact on the parameter of interest. However, this measure is only available for continuous data, whereas many important measures of latent variables in the social sciences are categorical. Examples include the votes (Yea/Nay) of senators indicating their ideology, the ranks (1–4) of respondents' value priorities, or the answers (correct/incorrect) to political knowledge questions. The following section therefore extends the EPC-interest to the case of categorical indicators.

2.2 *EPC-Interest*

The "EPC-interest" estimates the change in a free parameter estimate of the model that one can expect to observe if a particular restriction was freed. It is therefore a method of sensitivity analysis. However, the researcher is not forced to estimate all possible alternative models, but can evaluate the sensitivity of the results just by fitting the restrictive full invariance model. The EPC-interest is based on the work of Saris, Satorra, and Sörbom (1987), who introduced the expected parameter change in a fixed parameter for linear SEMs, and Bentler and Chou (1993), who introduced the expected parameter change in a free parameter after freeing a fixed parameter for SEM. It was applied to invariance testing with continuous data by Oberski (2014).

In models for categorical data, the EPC-interest can be derived, as shown in the online appendix, by applying general results of maximum-likelihood analysis. An additional problem with categorical data, however, is that there are usually sets of parameters relating to particular variables. For example, in Equation (2), there will be K "loadings" λ_{jk}, K "intercepts" τ_{jk}, K "direct effects" δ_{jk}, and, if present, K "interaction effects" δ^*_{jk} for each variable. This principle may be familiar from ANOVA: when an ANOVA term corresponds to a categorical variable, it will have several parameters that are considered simultaneously in the analysis. These parameters will, moreover, be strongly dependent on one another, so that the impact of freeing one of them cannot be seen separately from the impact of freeing a parameter relating to another category of the same variable.

For this reason, when extending EPC-interest to the categorical case, it also becomes necessary to allow for the consideration of *sets* of parameters to free rather than just investigating the effect of single restrictions. This requires a multivariate EPC-interest, rather than the univariate one in use so far.

Equation (2) shows that the hypothesis of measurement invariance often takes the form of restricting some parameter vector to zero. In Equation (2), this is the zero restriction on the "direct effect" $\delta_{jk} = 0$. But to accommodate categorical data models, we can consider a restriction on a vector of such parameters, $\boldsymbol{\Psi} = 0$. As shown in the online appendix, in its more general form the EPC-interest can then be written as

$$\text{EPC} - \text{interest} = \mathbf{P}\left(\frac{\partial\theta}{\partial\psi'}\right)(\psi - \hat{\psi}), \tag{3}$$

where $\mathbf{P}$ is a matrix selecting the parameters of interest, and $\boldsymbol{\theta}$ is the vector of free parameters of the model. That is, the EPC-interest can be seen simply as the coefficient of a linear approximation to the relationship between the free and fixed parameters, multiplied by the change in the fixed parameters. This demonstrates the difference with the sensitivity analysis approach common in econometrics (Magnus and Vasnev 2007, 168) in which only $\partial\boldsymbol{\theta}/\partial\boldsymbol{\Psi}'$ is considered: the EPC-interest combines both the direction ($\partial\boldsymbol{\theta}/\partial\boldsymbol{\Psi}'$) and the magnitude ($\boldsymbol{\Psi} - \hat{\boldsymbol{\Psi}}$) of the misspecification.

The accuracy of the approximation of the EPC-interest as a measure of the change in the parameters of interest is reflected in an order of approximation term, $O(\boldsymbol{\delta}'\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is the deviation from the true values $\boldsymbol{\delta} = \boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\Psi}')'$ (see the online appendix). This corresponds to results on the score test (modification index) and "EPC-self" in the literature on structural equation modeling, which can be shown to be exact under a "sequence of local alternatives"; that is, when $\boldsymbol{\vartheta} = \lim_{n\to\infty} \hat{\boldsymbol{\vartheta}} + n^{-\frac{1}{2}}\boldsymbol{\delta}$ (Satorra 1989, 135). It is important to note here that $\boldsymbol{\delta}$ is the deviation from the "true" value of $\boldsymbol{\vartheta}$, rather than the deviation from the limit of the parameter

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Table 1 Simulation study of EPC-interest

n	*250*			*500*			*1000*		
True δ	0	0.5	1	0	0.5	1	0	0.5	1
Est. $\hat{\gamma}$	1.010	1.151	1.353	0.980	1.152	1.330	1.013	1.163	1.327
Bias $\hat{\gamma}$	−0.010	−0.151	−0.353	0.020	−0.152	−0.330	−0.013	−0.163	−0.327
EPC-interest	0.003	−0.166	−0.494	−0.001	−0.180	−0.486	0.004	−0.182	−0.448

Note. Shown is the average point estimate for the γ parameter of interest under full measurement invariance (Est), its difference from the true value $\gamma = 1$ (Bias), and the average EPC-interest.

estimates under the alternative model. Therefore, another view on the accuracy is that it will be better when the alternative model is not strongly misspecified. For this reason it is important to consider freeing sets of very strongly related parameters simultaneously: if we considered freeing only one at a time, the alternative models would likely be misspecified.

2.3 *Small Simulation Study*

To demonstrate the extent of the approximation bias in the EPC-interest, we performed a small simulation study. In this study, we specified a latent variable model for four binary indicators: $P(Y_j = 1|x) = [1 + \exp(-x)]^{-1}$, with $j \in \{2, 3, 4\}$, and structural model $x = \gamma z + \epsilon$ with $\gamma = 1$ and $\epsilon \sim N(0, 1)$. We then introduced a violation of measurement invariance for the first indicator, $P(Y_1 = 1|x) = [1 + \exp(-x - \delta z)]^{-1}$. Nine conditions varied the sample size, $n \in \{250, 500, 1000\}$, and the size of the invariance violation: $\delta = 0$ (no violation), 0.5 (moderate), or 1 (extreme). Data were generated using R 3.1.2 and analyzed using Latent GOLD 5.0.0.14161.

The results over 200 samples are shown in Table 1. The first two rows show the average point estimate of the parameter of interest $\hat{\gamma}$, and its deviation from the true value $\gamma = 1$, respectively. It can be seen that a modest violation $\delta = 0.5$ still has a substantively important impact on bias. Under this condition, the amount of bias is reasonably well approximated by the EPC-interest, which has average values close to the true bias. When the violation is extreme, however, the approximation causes the EPC-interest to somewhat overestimate the bias: for example, in the last column of the table, the true bias is −0.327 but EPC-interest estimates it at about −0.448. The bias does not appear to be strongly affected by the sample size, demonstrating the asymptotic approximation results above.

Overall, the results of this very limited simulation study demonstrate the analytic results discussed above. As expected, for very large violations of invariance the EPC-interest will overestimate the bias caused somewhat. In these cases, the EPC-interest is still a useful guide but the researcher may wish to verify that after freeing the violation in question, the results of interest do indeed change substantially. Alternatively, where this is feasible, one may also resort to estimating all alternative models and examining the results. In other conditions, the EPC-interest performs as expected: when there is no bias, it estimates zero on average, and when there is moderate bias in the parameter of interest, EPC-interest approximates this bias adequately.

3 Example Application No. 1: 90th U.S. Senate Roll Call Data

To exemplify the use of the EPC-interest using a relatively simple latent variable model with categorical variables that is well known in political science, we estimate an ideal point model on roll call data for senators in the 90th U.S. Senate, which met from 1967 to 1969 during the Lyndon B. Johnson administration.

The probability that Senator i votes "Yea" on motion j is modeled as a logistic regression on the motion's (unobserved) utility,

$$P(\text{"Yea" on motion } j|x_i) = [1 + \exp(-\beta_j u_{ij})]^{-1}, \tag{4}$$

where the utility u_{ij} of the motion to that senator is simply the Euclidean distance between the

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

senator's position x_i and the motion's position τ_j,

$$u_{ij} = (x_i - \tau_j)^2. \tag{5}$$

This model is equivalent to the well-known unidimensional Poole and Rosenthal (1985) (W-NOMINATE) model. The latent value x_i is known as an "ideal point."

The Poole–Rosenthal model can be extended to incorporate covariates that predict the latent variable x_i, for example using the party of the senator as a predictor:

$$x_i = \alpha + \gamma \cdot \text{Party} + \epsilon_i. \tag{6}$$

Using party (Democratic or Republican) as a predictor allows the researcher to see how strongly party membership relates to senators' ideological positions, which ultimately influences their votes. A higher value of γ thus indicates more ideological homogeneity within parties in the Senate: we therefore call γ the "polarization coefficient."

The usual choice $\epsilon_i \sim \text{Normal}(0, \sigma^2)$ leads to a quadratic SEM for categorical data, or, equivalently, a quadratic 2PL IRT model with a covariate (Rabe-Hesketh, Skrondal, and Pickles 2004). Alternatively, the distribution of x can be estimated from the data by choosing some number K of categories for x (latent classes) and modeling the probability of belonging to class k as an ordered multinomial regression,

$$P(x_i = k|\text{Party}) = \frac{\exp(\alpha_k + \gamma \cdot x^{(k)} \cdot \text{Party})}{\sum_{m=1}^{K} \exp(\alpha_m + \gamma \cdot x^{(m)} \cdot \text{Party})}, \tag{7}$$

where $x^{(k)}$ is a latent score assigned to the kth category of x. For this arbitrary choice of latent scale, we choose $x^{(k)}$ to go from 0 to 1 in equally spaced intervals (following Vermunt and Magidson 2013). The latent category intercepts α_k allow the distribution of the latent dimension to be freely estimated rather than assumed normal. This leads to the "latent class factor model" (Vermunt and Magidson 2004).

A possible problem when using ideal point models to investigate polarization is that it is assumed that this polarization is the same on all motions the Senate votes on. If there is some motion on which the votes of senators from the same party are significantly more tightly knitted than usual, there will be an effect of party over and above that of the senator's utility for this motion. A vote model with a direct covariate effect,

$$P(\text{"Yea" on motion } j|x_i) = [1 + \exp(-\beta_j u_{ij} - \delta_j \cdot \text{Party})]^{-1}, \tag{8}$$

then replaces Equation (4). In other words, the usual ideal point model assumes measurement invariance with respect to the covariate, an assumption that can be expressed as $\delta_j = 0$. Where such direct effects do exist, they are relevant to the investigation of polarization insofar as ignoring them biases the estimate of the Senate's overall polarization. Thus, we investigate whether the assumption of measurement invariance $\delta_j = 0$ seriously affects the estimate of interest $\hat{\gamma}$ using the EPC-interest.

Maximum-likelihood estimates of the parameters were obtained using the software Latent GOLD, taking $K = 4$ and the first twenty motions introduced in the 90th Senate as an example. The model appears to fit the data well, with an L^2 bootstrapped p-value of 0.14, The factor scores $\hat{x}_i$ obtained from this simple model correlated highly (0.79) with those obtained from W-NOMINATE (Poole et al. 2011) and from Optimal Classification Roll Call Scaling (0.81; Poole et al. 2012). Based on the full invariance model, the polarization coefficient $\hat{\gamma}$ was estimated at 4.164 (SE 1.3077). Since γ is a logistic regression coefficient, a Republican senator has about four times higher odds of being one category above a Democratic senator than of being in the same category.[3] There was therefore considerable polarization in the 90th Senate.

[3] Because there are four x categories scored $\{0, 1/3, 2/3, 1\}$, the odds are $\exp(\hat{\gamma}/3) \approx 4$.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

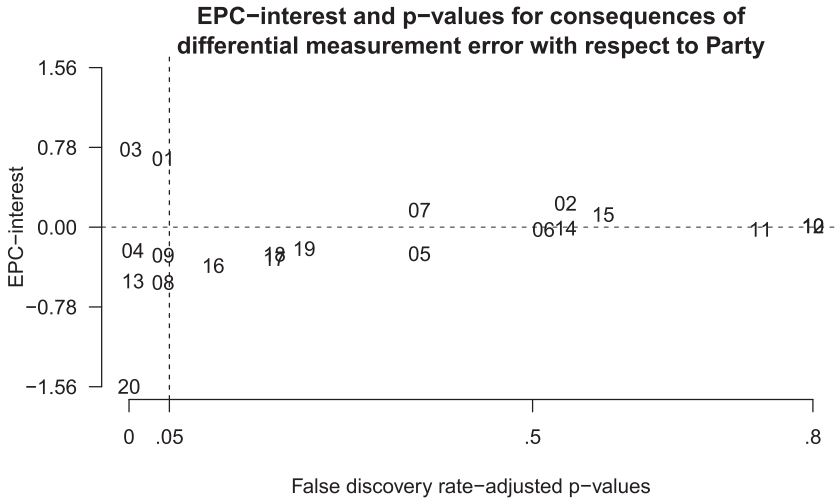


Fig. 2 The effect of twenty measurement invariance assumptions on the polarization parameter of interest γ, plotted against the p-values for each violation.

However, violations of measurement invariance could conceivably bias this conclusion. To investigate this, Fig. 2 plots the EPC-interest values of freeing the direct effects δ_j on the parameter estimate of interest $\hat{\gamma}$. Each number in Fig. 2 corresponds to a motion number introduced in the Senate. The vertical axis, labeled "EPC-interest," estimates the change from the current estimate ($\hat{\gamma} = 4.164$) under full measurement invariance that one can expect to observe in $\hat{\gamma}$ when freeing the direct effect of party for that motion. The horizontal axis shows the p-value for the null hypothesis that the corresponding $\delta_j = 0$, adjusted for false discovery rate (Benjamini and Hochberg 1995). The idea behind plotting both of these quantities at the same time is that the researcher will likely be interested in violations of measurement invariance that are both statistically and substantively significant (Saris, Satorra, and Van der Veld 2009).

Figure 2 shows that, of the statistically significant violations of measurement invariance, motion no. 03 violates measurement invariance in a manner that augments the estimated polarization (EPC-interest is positive). Motions no. 04 and no. 13 violate measurement invariance in an approximately opposite manner (EPC-interests are negative). However, motion no. 20 clearly stands out as an important violation of measurement invariance. After introducing the direct effect of party on voting "Yea" to motion no. 20 (freeing δ_{20}), no other measurement invariance violations are statistically significant (all false discovery rate-adjusted p-values ≥ 0.05). Thus, it seems that motion no. 20 (HR4573, which increased the public debt limit) was an issue on which the ranks were closed more than usual. Indeed, the 1967 CQ Almanac[4] specifically reports on this motion, remarking that "Republicans launched their first major attacks in the 90th Congress on the Administration's fiscal policies," with no Republicans voting in favor.

After adjusting for this event, the polarization coefficient is estimated at 3.422 (SE 1.0630): the tight-knittedness between party membership and voting pattern is therefore somewhat loosened, but still strong. The partial invariance model after accounting for this one violation becomes acceptable, in the sense that those violations that are present do not substantially change the results of interest regarding polarization. The model fit the data well with an L^2 bootstrapped p-value of 0.11. Its BIC (1379) indicates an improvement over that of the full invariance model (1405). Overall, the difference between the fully and partially invariant models is modest. Figure 3 shows the effect of freeing δ_{20} on the "ideal point" estimates; that is, the latent variable estimates $\hat{x}_i$.

The top part of Fig. 3 shows the ideal point estimates under the full invariance model. These estimates range from zero to one; since Democrats, shown as black dots, are predominantly on the

[4]http://library.cqpress.com/cqalmanac/document.php?id=cqal67-1314297.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

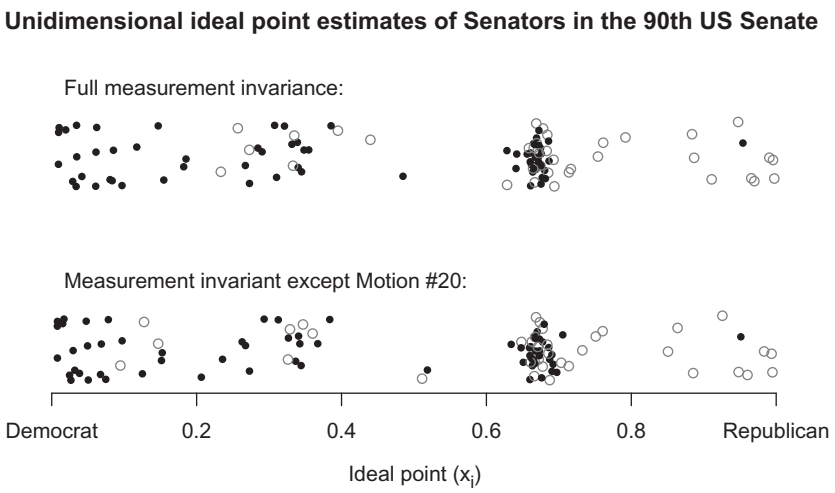


Fig. 3 Posterior point estimates of the latent variable "ideal point" score $\hat{x}_i$ for all senators (points), jittered vertically for visual clarity. Filled points are Democrats, open circles Republicans. Top: estimates under the fully invariant model; Bottom: estimates allowing for the direct effect $\delta_{20} \neq 0$.

lower side of the scale, zero has been labeled "Democrat" and the score 1 has been labeled "Republican," since most Republicans (open circles) can be found here. To make the points more visible, they have been jittered randomly in the vertical direction. The bottom part of Fig. 3 shows the ideal point estimates for the same senators, but this time while accounting for the partial violation of measurement invariance $\delta_{20} \neq 0$. It can be seen that Republicans are more spread out into the "Democratic" side of the scale. Especially the three Republican senators with a score below 0.2 experience a rather large shift in position. On the whole, therefore, the differences between the two distributions are modest, but the differences for individual senators' ideal point estimates can be quite substantial.

In this section, we investigated measurement invariance assumptions in the well-known "ideal point" model for binary roll call data. The example demonstrated that even when the model fits the data well initially, it is still possible for violations of measurement invariance to bias the conclusions. The EPC-interest for categorical data was used here as a tool to detect such bias. After accounting for one violation of measurement invariance, the final model differed somewhat from the original conclusions: the estimated amount of polarization in the 90th Senate was lower, and several Republican senators' estimated ideological positions were considerably more liberal.

4 Example Application No. 2: Ranking Values in the WVS

Our second, more complex, example application employs the 2010–2012 World Values Survey[5] (WVS) comprising $n = 67{,}568$ respondents in forty-eight different countries (the online appendix provides a full list of countries). The WVS questionnaire includes Inglehart's (1981) extended (post)materialism questions, developed to measure political values priorities. This extended version includes three sets of four priorities (Table 2) to be ranked by the respondents. Of these, set B in Table 2 is known as the "short scale" that is commonly used in research on values priorities.

Based on the "dual hypothesis model" (Inglehart 1981, 881), previous authors have suggested a structural relationship of interest between, on the one hand, (post)materialism, and socio-economic (Inglehart and Welzel 2010) as well as socio-cultural (Inglehart, Norris, and Welzel 2002) variables, on the other. We will follow these authors and examine the aggregate relationship of values priorities with log-GDP per capita (Z_1) and the percentage of women in parliament (Z_2).[6]

[5]http://www.worldvaluessurvey.org/.
[6]These country-level variables were obtained from the World Bank database[7] using the WDI package (Arel-Bundock 2013) in R 3.0.2 (R Core Team 2012).
[7]http://data.worldbank.org/.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Table 2 Value priorities to be ranked for the three WVS 2010–2012 ranking sets

Option number	*M*/*P*	*Value wording*
Set A		
1.	M	A high level of economic growth
2.	M	Making sure this country has strong defense forces
3.	P	Seeing that people have more say about how things are done at their jobs and in their communities
4.	P	Trying to make our cities and countryside more beautiful
Set B		
1.	M	Maintaining order in the nation
2.	P	Giving people more say in important government decisions
3.	M	Fighting rising prices
4.	P	Protecting freedom of speech
Set C		
1.	M	A stable economy
2.	P	Progress toward a less impersonal and more humane society
3.	P	Progress toward a society in which ideas count more than money
4.	M	The fight against crime

Notes. "Materialist" concerns are marked "M," "postmaterialist" concerns are marked "P." Wording: "People sometimes talk about what the aims of this country should be for the next 10 years. On this card are listed some of the goals which different people would give top priority. Would you please say which one of these you, yourself, consider the most important?"; "And which would be the next most important?"

We model the probability that unit i in country c belongs to category x of a latent (post)materialism variable with T classes using the multilevel multinomial logistic regression

$$P(X_{ic} = x|Z_{1ic} = z_{1ic}, Z_{2ic} = z_2, G_c = g) = \frac{\exp(\alpha_x + \gamma_{1x}z_1 + \gamma_{2x}z_2 + \beta_{gx})}{\sum_t \exp(\alpha_t + \gamma_{1t}z_1 + \gamma_{2t}z_2 + +\beta_{tg})}, \tag{9}$$

where the country-level random effect variable G has been introduced. We take the "random effects" variable G to be a country-level latent class variable with S classes and a freely estimated (nonparametric) distribution. Overall, then, our model can be seen as a multilevel multinomial regression of (post)materialism on country-level covariates, in which the nominal dependent variable is latent, and the random effects distribution is nonparametric. The main parameters of substantive interest in Equation (9) are therefore the multinomial logistic regression coefficients γ_{mx}. This "structural" part of the model is shown in the top part of Fig. 4.

The latent (post)materialism variable is measured by respondents' rankings of value priorities. Each respondent in the forty-eight countries has ranked only their first and second choice on three ranking tasks A, B, and C (see Table 2). We assume that each value priority has a particular "utility" (Luce 1959; Böckenholt 2002, 171) dependent on the latent class variable (Croon 1989; Böckenholt 2002, 172). For example, for ranking task A in country c, respondent i's choices for first and second place are modeled as

$$P(A_{1ic} = a_1, A_{2ic} = a_2|X_{ic} = x) = \frac{\exp(\tau_{a_1} + \lambda_{a_1x})}{\sum_k \exp(\tau_k + \lambda_{kx})} \frac{\exp(\tau_{a_2} + \lambda_{a_2x})}{\sum_{k \neq a_1} \exp(\tau_k + \lambda_{kx})}. \tag{10}$$

Crucially, the choice for second place, A_{2ic}, is modeled while excluding the alternative already chosen for first place, A_{1ic}. In other words, we model a sequential ranking process in which first place is chosen from all options, then second place is chosen from the remaining options. This dependency intrinsic to ranking tasks makes the model used here different from standard models for categorical dependent variables. The measurement model for (post)materialism is shown in the bottom part of Fig. 4, using gray arrows to indicate that the choice for second place is modeled conditionally upon that for first place.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

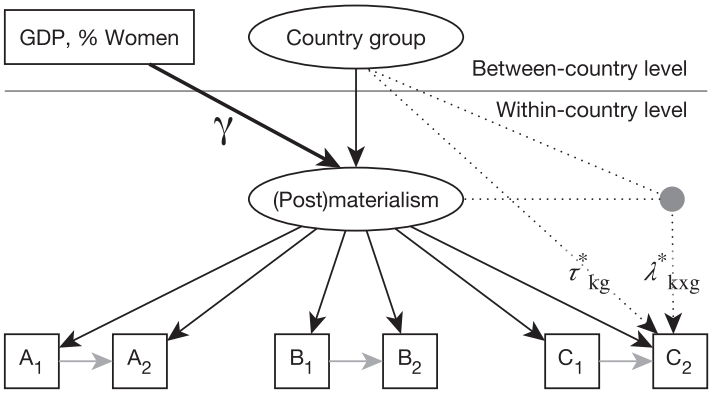


Fig. 4 Graphical representation of the multilevel latent class regression model for (post)materialism measured by three partial ranking tasks. Observed variables are shown in rectangles, while unobserved (latent) variables are shown in ellipses.

Measurement invariance violations could be parameterized by extending Equation (10) to include country group class (g) effects:

$$P(A_{1ic} = a_1, A_{2ic} = a_2 | X_{ic} = x) = \frac{\exp(\tau_{a_1} + \lambda_{a_1 x} + \delta_{a_1 g} + \delta^*_{a_1 xg})}{\sum_k \exp(\tau_k + \lambda_{kx} + \delta_{kg} + \delta^*_{kxg})} \ldots \tag{11}$$

The base invariance model in Equation (10) can be seen as fixing the direct main effects $\delta_{kg} = 0$ and direct interaction effects $\delta^*_{kxg} = 0$. Figure 4 shows an example for the second ranking of Set C (C_2) as the dotted main effect and interaction effects. In line with Section 2, we now investigate whether the possible misspecifications in the measurement invariance model, $\delta_{kg} \neq 0$ and $\delta^*_{kxg} \neq 0$, substantially affect the parameters of interest γ_{mx} using EPC-interest for categorical data.

The full measurement invariance model including parameters of interest γ_{mx} was estimated using Latent GOLD 5. Following Moors and Vermunt (2007), we select three classes for both the latent "(post)materialism" variable and the country group class variable, $T = S = 3$. For BIC values and the rationale behind these choices, please see the online appendix. Since class selection is not the focus of this example analysis, we will not discuss it here further.

After estimating the full invariance model, we calculated the EPC-interest for the δ and δ^* parameters, of which our three-class model has four: two for each of the two independent variables. Measurement invariance violations can potentially take the form of six direct main effects (δ_{jkg}) and twelve direct interaction effects (δ^*_{jkxg}) for each of the three ranking tasks, totaling fifty-four possible misspecifications in the full invariance model. These misspecifications are strongly correlated and should not be considered separately. Rather, we consider the probable impact of freeing the direct main effects for each ranking task separately and of freeing the direct interactions for each ranking task separately. Thus, rather than considering the direct and interaction effects for each of the forty-eight countries on each of the three unique categories of each of the three ranking tasks, making 5076 potential EPC-interest values, we evaluate direct effects of the country group random effect and consider their impact jointly for strongly correlated misspecifications, reducing the problem to twenty-four EPC-interest values of interest.

Table 3 shows these twenty-four EPC-interest values together with the parameter estimates from the full invariance model. The EPC-interest values estimate the change from the current estimates of interest after freeing the direct main effects (δ_{jkg}) or interaction effects (δ^*_{jkg}). In the full invariance model, Class 1 corresponds to a "postmaterialism" class. The estimate −0.035 (SE 0.007) shown in Table 3 would therefore suggest that more prosperous nations tend to be less postmaterialist. This directly contradicts the theory of Inglehart and Welzel (2010).

Since the theory specifies only that certain coefficients should be positive or negative, the key focus of substantive interest is whether misspecifications can potentially change a parameter of interest's sign. In Table 3, we therefore look for EPC-interest values that would change the sign of

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Table 3 Full invariance multilevel latent class model: Parameter estimates of interest with standard errors, and EPC-interest when freeing each of six sets of possible misspecifications

				EPC-interest for...					
				δ_{jkg}			δ^*_{jkxg}		
		Estimates $\widehat{\gamma_{mx}}$		*Ranking task*			*Ranking task*		
		Est.	*SE*	*1*	*2*	*3*	*1*	*2*	*3*
Class 1	GDP	−0.035	(0.007)	−0.013	0.021	−0.002	**0.073**	**0.252**	0.005
Class 2	GDP	−0.198	(0.012)	−0.018	−0.035	0.015	−0.163	−0.058	0.002
Class 1	Women	0.013	(0.001)	−0.006	0.002	0.000	−0.003	0.029	0.002
Class 2	Women	−0.037	(0.001)	0.007	−0.003	0.002	−0.006	−0.013	0.002

those estimates. These are shown in bold face in Table 3. Two such EPC-interest values are indeed present, namely the direct interaction effect of the country group class with the postmaterialism class on ranking tasks 1 and 2. This means that the attribute parameters that define the classes for these two tasks differ over country groups, and that after accounting for these differences the effect of GDP on postmaterialism is estimated to be positive rather than negative. This set of misspecifications is thus of substantive interest and should be amended in the model.

Following the common practice of partial invariance models, we free these two sets of measurement invariance violations, allowing for differences in the parameters of ranking tasks 1 and 2 across country groups. Table 4 shows the substantive parameter estimates and EPC-interest values for the resulting partial invariance model. The substantive regression coefficient for the effect of GDP on the postmaterialism class (Class 2 in Table 4) is indeed positive after freeing the detected misspecifications. Recalculating EPC-interest values for the remaining possible misspecifications reveals that none of the possible misspecifications in this partial invariance model has the potential to change the substantive conclusions. We therefore conclude that the partial invariance model fits "approximately," since none of the substantive conclusions based on it are threatened by measurement invariance violations.

This section demonstrated the use of the EPC-interest for measurement invariance testing in a more complex example. A violation of measurement invariance was detected that could reverse the conclusions of substantive interest. After accounting for this violation, no further such violations are detected.

5 Discussion and Conclusion

Whenever groups are compared, measurement invariance is a concern. Particularly, it should be verified that substantive conclusions of interest are uncontaminated by possible cross-group differences in measurement. The "expected parameter change in the parameter of interest" or EPC-interest, a measure introduced by Oberski (2014) for this purpose in the context of linear SEMs, was extended in this article to categorical observed and latent variables as well as rankings and other types of data often encountered in the social sciences.

The EPC-interest for categorical data is an approximation of the change in the parameters of substantive interest that we can expect to observe if a particular violation of measurement invariance was freed. A small simulation study showed that this approximation works well when the misspecification is moderate, and overestimates the bias somewhat when it is extreme. In this case, EPC-interest still indicates the most important violations of measurement invariance but the researcher may wish to verify that the expected parameter change is close to the actually observed change.

Two example applications of categorical data latent variable models demonstrated the utility of the EPC-interest. The first modeled U.S. Senators' latent ideology as a function of party

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Table 4 Partially invariant multilevel latent class model: Parameter estimates of interest with standard errors, and EPC-interest when freeing each of six sets of possible misspecifications

				EPC-interest for non-invariance of...				
				δ_{kg}			δ^*_{kxg}	
		Estimates $\hat{\gamma}_{mx}$		*Ranking task*			*Ranking task*	
		Est.	*SE*	*1*	*2*	*3*	*1*	*3*
Class 1	GDP	−0.127	(0.008)	−0.015	−0.003	0.002		0.097
Class 2	GDP	0.057	(0.011)	−0.043	−0.013	0.002		0.161
Class 1	Women	0.008	(0.001)	−0.002	0.000	0.002		0.001
Class 2	Women	0.020	(0.001)	−0.007	−0.001	0.002		0.007

membership to estimate polarization. After fitting the full invariance model, EPC-interest detected one violation of measurement invariance that substantially reduced the estimated polarization and made the ideal point estimates for some Republican senators considerably more liberal. The second example application looked at the relationship between latent (post)materialism on the one hand and, on the other hand, log-GDP per capita and the percentage of women in parliament using data from 67,568 respondents in forty-eight countries. Violations of measurement invariance existed that, when unaccounted for, could reverse substantive conclusions. The EPC-interest allowed us to detect this problem and prevent it.

We suggest using the EPC-interest or other sensitivity analyses to evaluate measurement invariance as we did in the examples discussed above: start from a model that assumes measurement invariance, and evaluate the impact of these restrictions on the conclusions of interest. The entire resulting sequence of models should be reported so readers can reproduce the results and decide whether they agree that violations removed could be judged as substantively "large" and those left in the model as "small." An example of such a judgment, given in the application, is that a positive effect can be changed to a negative one or vice versa; in other applications, different criteria will be relevant.

Is it appropriate to free violations of measurement invariance where encountered?[8] Freeing violations has long been standard practice in both the "partial invariance" literature (Byrne, Shavelson, and Muthén 1989) and the DIF literature (Holland and Wainer 1993), but there are situations in which it can be misleading—especially when the alternative model cannot identify further free parameters. An example is a model with only one invariance restriction so that the alternative model is simply the model with only *a priori* reference indicators. In such cases the choice of indicator for which to free the parameters (i.e., the choice of reference indicator) cannot be determined based on model fit, and will often lead to opposite changes in the parameters of interest (Hancock, Stapleton, and Arnold-Berkovits 2009). The EPC-interest can then still be used to investigate whether the conclusions are robust to violations of the assumptions. When they are not, however, there is no empirical basis on which to select a model. In other situations, such as in the examples, however, the alternative model still imposes testable restrictions on the data. It then seems reasonable to proceed by freeing some restrictions while retaining others, as is the common practice.

Funding

The Netherlands Organization for Scientific Research (NWO) (Veni grant number 451-14-017) to D.L.O. and (Vici grant number 453-10-002) to J.K.V.

[8]We thank an anonymous reviewer for raising this question.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Conflict of interest statement. None declared.

References

Arel-Bundock, Vincent 2013. WDI: World Development Indicators (World Bank). R package version 2.4. http://CRAN.R-project.org/package=WDI (accessed August 1, 2014).

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.

Bentler, P. M., and C. P. Chou. 1993. Some new covariance structure model improvement statistics. In *Testing structural equation models*, eds. K. G. Jöreskog and J. Scott Long, 199–235. Thousand Oaks, CA: Sage.

Böckenholt, U. 2002. Comparison and choice: Analyzing discrete preference data by latent class scaling models. In *Applied latent class analysis*, eds. Jacques A. P. Hagenaars and Allan L. McCutcheon, 163–82. Cambridge, UK: Cambridge University Press.

Bollen, Kenneth A. 2002. Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53:605–34.

Byrne, B. M., R. J. Shavelson, and Bengt Muthén 1989. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105:456.

Chen, F. F. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling* 14:464–504.

Cheung, G. W., and R. B. Rensvold. 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling* 9:233–55.

Croon, Marcel 1989. Latent class models for the analysis of rankings. In *New developments in psychological choice modelling*, eds. G. De Soete, H. Feger, and K. C. Klauer, 99–121. North-Holland: Elsevier Science Publishers.

Hancock, Gregory R., Laura M. Stapleton, and Ilona Arnold-Berkovits. 2009. The tenuousness of invariance tests within multisample covariance and mean structure models. In *Structural equation modeling in educational research: Concepts and applications*, eds. T. Teo and M. S. Khine, 137–74. Rotterdam, The Netherlands: Sense Publishers.

Holland, Paul W., and Howard Wainer. 1993. *Differential item functioning*. New York: Routledge.

Hu, L., and P. M. Bentler. 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods* 3:424.

Inglehart, Ronald. 1981. Post-materialism in an environment of insecurity. *American Political Science Review* 75:880–900.

Inglehart, Ronald, and Christian Welzel. 2010. Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics* 8:551–67.

Inglehart, Ronald, Pippa Norris, Christian Welzel. 2002. Gender equality and democracy. *Comparative Sociology* 1:321–45.

Kankaraš, Miloš, Guy Moors, and Jeroen K. Vermunt. 2010. Testing for measurement invariance with latent class analysis. In *Cross-cultural analysis: Methods and applications*, eds. Eldad Davidov, Peter Schmidt, and Jaak Billiet, 359–84. New York: Taylor & Francis.

Luce, R. Duncan 1959. *Individual choice behavior: A theoretical analysis*. New York: John Wiley and Sons.

Magnus, J. R., and A. L. Vasnev. 2007. Local sensitivity and diagnostic tests. *Econometrics Journal* 10:166–92.

McCutcheon, Allan L. 1985. A latent class analysis of tolerance for nonconformity in the American public. *Public Opinion Quarterly* 49:474–88.

Mellenbergh, G. J. 1989. Item bias and item response theory. *International Journal of Educational Research* 13:127–43.

Meredith, W. 1993. Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58:525–43.

Moors, Guy, and Jeroen Vermunt. 2007. Heterogeneity in post-materialist value priorities: Evidence from a latent class discrete choice approach. *European Sociological Review* 23:631–48.

Oberski, Daniel 2015. Replication Data for: "Evaluating measurement invariance in categorical data latent variable models with the EPC-interest." http://dx.doi.org/10.7910/DVN/I7Y3G2 (accessed July 1, 2015).

Oberski, D. L. 2014. Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis* 22:45–60.

Poole, Keith, Jeffrey Lewis, James Lo, and Royce Carroll. 2012. *oc: OC Roll Call Analysis Software*. http://cran.r-project.org/web/packages/oc/index.html (accessed August 1, 2014).

Poole, Keith T., and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science* 29(2):357–84.

Poole, Keith T., Jeffrey B. Lewis, James Lo, and Royce Carroll. 2011. Scaling roll call votes with w-nominate in R. *Journal of Statistical Software* 42(14).

R Core Team. 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/ (accessed August 1, 2014).

Rabe-Hesketh, Sophia, Andrew Skrondal, and Andrew Pickles. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69:167–90.

Saris, W. E., A. Satorra, and D. Sörbom. 1987. The detection and correction of specification errors in structural equation models. *Sociological Methodology* 17:105–29.

Saris, W. E., A. Satorra, and W. M. Van der Veld. 2009. Testing structural equation models or detection of misspecifications? *Structural Equation Modeling* 16:561–82.

Satorra, A. 1989. Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika* 54:131–51.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press

Schmitt, N., and G. Kuljanin. 2008. Measurement invariance: Review of practice and implications. *Human Resource Management Review*18:210–22.

Steenkamp, J. B. E. M., and H. Baumgartner. 1998. Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25:78–107.

Vandenberg, R. J., and C. E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3:4–70.

Vermunt, J. K., and J. Magidson 2004. Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In *New developments in categorical data analysis for the social and behavioral sciences*, eds. L. Andries van der Ark, Marcel A. Croon, and Klaas Sijtsma, 41–63. Mahwah: Erblaum.

Vermunt, J. K., and J. Magidson. 2013. *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

https://doi.org/10.1093/pan/mpv020 Published online by Cambridge University Press