Psychological Test and Assessment Modeling, Volume 59, 2017 (3), 317-341

Measurement models for ordered-categorical indicators: A factor analytic approach for testing the level of measurement

Takuya Yanagida[1], *Petra Gradinger*[2] *& Dagmar Strohmeier*[2]

Abstract

In the social sciences, self-reports administered as questionnaires are frequently used to measure psychological constructs. Data stemming from scale items are commonly analyzed using statistical methods for metric dependent variables. However, the assumption of interval level data is not tested but assumed to be fulfilled. One reason for ignoring this assumptions is the lack of an adequate approach for testing this assumption. Thus, we present a factor analytic approach for testing the level of measurement. First, two empirical examples are presented to demonstrate this approach. Second, a simulation study based on several conditions with varying population model and sample size was conducted. Results of the simulation study demonstrate the functioning of this approach. In sum, the factor analytic approach can be used for testing the level of measurement of scale items enabling empirical decision making about choosing appropriate statistical methods instead of relying on untested assumptions.

Key words: confirmatory factor analysis, measurement model, level of measurement, ordinal data, statistical assumption

[1] *Correspondence concerning this article should be addressed to:* Takuya Yanagida, PhD, Department of Applied Psychology: Work, Education, and Economy, Faculty of Psychology, University of Vienna, Universitätsstraße 7 (NIG), 1010 Vienna, Austria; email: takuya.yanagida@univie.ac.at

[2] University of Applied Sciences Upper Austria, Austria

In the social and behavioral sciences, self-report measures administered as questionnaires are commonly used to measure psychological constructs. Data resulting from these measurements are subjected to statistical methods to describe characteristics of the sample (i.e., descriptive statistics) and to draw conclusions about the population (i.e., inferential statistics). Thus, appropriate statistical methods need to be selected based on research hypotheses and in accordance with the data at hand. One of the fundamental deciding factors for choosing appropriate statistical methods is the level of measurement of data (Pett, 2015). However, researchers rarely bother about the level of measurement and oftentimes apply statistical tests for metric dependent variables (e.g., two-sample *t*-test or ANOVA) instead of statistical tests commonly used for ordered-categorical dependent variables (e.g., Mann-Whitney U test or Kruskal-Wallis test). One reason for simply assuming metric-level data is that there is no appropriate statistical method available to test this assumption.

In the present study, we introduce a factor analytic approach for testing the level of measurement based on measurement models for ordered-categorical indicators. In the following, we present a commonly used item format for questionnaires and discuss the concept of level of measurement. Next, we discuss two approaches for analyzing measurement models for ordered-categorical indicators. Last, we describe the analytical steps for testing the level of measurement.

Note that the present paper focuses on questionnaires items. As for items in achievement or intelligence tests, the dichotomous Rasch model and the uni- or multidimensional polytomous Rasch model (Rasch, 1960/1980; see also Fischer, 1974) can be used to test hypotheses about the properties of the items (see Hohensinn & Kubinger, 2017).

Questionnaire items

A questionnaire item consists of an item stem, which contains the stimulus material to which respondents have to respond, and a system of response options (McDonald, 1999). There are numerous item types from which rating scale items are most frequently used in the social and behavior sciences. In a rating scale item, respondents are asked to answer a specific question (e.g., How much do you enjoy scientific writing?) by selecting an option out of a set of ordered categories (e.g., *not at all – very little – somewhat – to a great extent*). Numerical values are assigned to the item responses in accordance to the selected response options. The amount of information within the numbers assigned to the item responses is described in the level of measurement.

Level of measurement

According to Stevens (1960), four levels of measurement are distinguished: (1) nominal, (2) ordinal, (3) interval, and (4) ratio. The former two are labeled categorical and the latter two metric. More specifically, a categorical variable (see Agresti, 2002) comprises a set of categories which can be either unordered (i.e., numbers represent a qualitative distinction, e.g., girls and boys) or ordered (i.e., numbers represent some natural order,

e.g., school grades). A metric variable, on the other hand, not only represents a meaningful order of numbers, but also meaningful intervals between numbers (i.e., number represent equal differences in the underlying variable, e.g., temperature in degrees Fahrenheit). In theory, the distinction between ordered-categorical and metric variables is clear. In practice, however, it is not clear if data resulting from scale items of a questionnaire are ordered-categorical or metric. As for rating scale items, it seems clear that the resulting data is ordered-categorical and not metric given ordered categories of response options. Nevertheless, scores on rating scale items are typically summed to derive a composite index of the construct of interest used in subsequent statistical analyses. The implicit assumption of this approach is that respondents perceive differences between adjacent levels of response categories as equidistant. For example, the equal-distance coding *not at all* = 0, *very little* = 1, *somewhat* = 2, and *to a great extent* = 3 assumes that the distance between *not at all* and *very little* represents the same amount of the construct as the distance between *somewhat* and *to a great extent*.

In sum, researchers rarely question metric level of data resulting from a questionnaire based on rating scale items. Consequently, data are subjected to statistical methods which in fact are designed to analyze metric data. Theses analyses might result in biased parameter estimates and wrong conclusions about the research hypotheses. For that reason, ignoring the level of measurement is known to be a measurement error fallacy that is most consequential and prevalent in published quantitative research (see Wang, Watts, Anderson & Little, 2013).

One reason for ignoring the level of measurement might be that there is no appropriate statistical method available to test this assumption. Thus, the goal of the present study is to present a factor analytic approach for testing the level of measurement.

Factor analytic approach for testing the level of measurement

The factor analytic approach for testing the level of measurement is based on a measurement model for ordered-categorical indicators (Bovaird & Koziol, 2012). There are two approaches for analyzing these models: (1) the *underlying response variable approach* and (2) the *response function approach* (Jöreskog & Moustaki, 2001).

Underlying response variable approach. This approach assumes that underlying each categorical scale item $\boldsymbol{y}_j$ is a continuous and normally distributed latent response variable $\boldsymbol{y}_j^*$ that leads to an observed ordinal variable (Muthén, 1983). That is, the observed ordinal variables are regarded as a crude measurement of underlying unobserved continuous variables. It is assumed that the observed ordinal response y_{ij} of respondents i (i = 1, 2, …, n), where n equals the sample size, on item j (j = 1, 2, …, m), where m equals the number of items, is related to the latent response y_{ij}^* via a threshold model so that

$$y_{ij} = \begin{cases} 0 & \text{if} \quad -\infty < y_{ij}^{*} \le \tau_{1j} \\ 1 & \text{if} \quad \tau_{1j} < y_{ij}^{*} \le \tau_{2j} \\ \vdots & \qquad\qquad \vdots \\ c-1 & \text{if} \quad \tau_{(c-2)j} < y_{ij}^{*} \le \tau_{(c-1)j} \\ c & \text{if} \quad \tau_{(c-1)j} < y_{ij}^{*} \le \infty \end{cases}$$

where c denotes the number of response categories and τ_{kj} denotes the threshold parameter of the threshold k (k = 1, 2, …, c – 1) of item j, which represent the location of the cut points along the latent response variable $\boldsymbol{y}^{*}$ (Skrondal & Rabe-Hesketh, 2005). That is, the threshold represents the critical point where respondents transition from one response category to another. Note that there are always c –1 thresholds involved.

Model estimation for the measurement model is based on a limited-information approach (Edwards, Wirth, Houts & Xi, 2012), which uses only a summary of the available data (i.e., polychoric correlations). The most common limited information method is the robust weighted least square estimator (DWLS) using a diagonal weight matrix (see Muthén, du Toit & Spisic, 1997).

Response function approach. This approach directly models the nonlinear relationship between the observed ordinal response y_{ij} of a respondent i on item j and a normally distributed latent factor $\boldsymbol{f}$ using a generalized linear model (Skrondal & Rabe-Hesketh, 2004).

The factor analytic model for polytomous items is

$$P\left(y_{ij} = r | f_i, \lambda_j, \tau_{kj}\right) = \frac{1}{1+e^{\left(\tau_{kj} - \lambda_j f_i\right)}} - \frac{1}{1+e^{\left(\tau_{(k+1)j} - \lambda_j f_i\right)}} = P\left(y_{ij} \ge r\right) - P\left(y_{ij} \ge r+1\right)$$

According to this model, the probability that the response y_{ij} of a respondent i to a given item j equals a specific response option category r (r = 1, 2, …, c) is a function of f_i, the individual i's location on the continuum of the latent construct $\boldsymbol{f}$. The factor loading λ_j reflects the strength of association between the latent factor $\boldsymbol{f}$ and the item j. The threshold parameter τ_{kj} of threshold k (k = 1, 2, …, c – 1) for item j indicates the extent to which the individual must possess the latent construct in order to transition from a lower response category to the next higher category at a chance level of 50%. Again, there are c – 1 threshold parameter, where c denotes the number of categories.

Model estimation for the measurement model using confirmatory factor analysis is based on a full-information approach (Edwards et al., 2012), which uses the raw data rather than summary statistics like in the response variable approach. Note that the chi-square test and fit indices (see West, Taylor & Wu, 2012) are not available using the response function approach using a full-information approach.

Analytical steps for testing the level of measurement

The factor analytic approach for testing the level of measurement comprises two consecutive steps and is based on the *response function approach* of measurement models for ordered-categorical indicators. That is, all measurement models are estimated using confirmatory factor analysis based on marginal maximum likelihood (MML) estimation with numerical integration. The *response function approach* was chosen because information criteria are needed to evaluate relative fit of competing models, which are only available with a full-information approach. Moreover, a full-information approach is known to be a more efficient estimator than a limited-information approach (see Lei & Qu, 2012).

Step 1: Estimate measurement models. In the first step, different measurement models for ordered-categorical indicators are estimated. These models differ in the assumptions about the level of measurement of the indicators, which are specified using parameter constraints regarding the location of threshold parameters. Thresholds of the (1) *interval scale model* are constraint to be equally spaced across all items, whereas the thresholds of the (2) *ordinal scale model* are freely estimated without any constraints. These models are equivalent to the theoretical differences between ordered-categorical and metric scale items (Andrich, 1982). Another measurement model, which accounts for the ordinal nature of scale items, but is more parsimonious than the *ordinal scale model* is the *rating scale model* (Andrich, 1978). Thresholds of the (3) *rating scale model* are not constrained to be equally spaced, but are constrained to have the same unequal spacing across all items.

Step 2: Decide between competing models. In the second step, one has to decide between competing models by comparing the *interval scale model vs. ordinal scale model*. Model comparison in favor of the *ordinal scale model* indicates that the scale items are not interval scale but ordered-categorical. In case the *ordinal scale model* was chosen, an additional model comparison can be employed comparing the *rating scale model* vs. the *ordinal scale model*. However, the first model comparison is sufficient to test whether scale items are ordered-categorical or metric.

Satorra-Bentler (S-B) scaled chi-square difference test and information criteria (AIC and BIC) can be used for model comparison. A statistically significant chi-square difference test for the comparison *interval scale model* vs. *ordinal scale model* is in favor of the *ordinal scale model*. A lower AIC and BIC value indicates a better trade-off between model fit (i.e., model deviance) and model complexity (i.e., number of estimated parameters). In addition, the location of threshold parameters of all items based on the *ordinal scale model* should be graphically inspected to aid decision making between competing models.

Present study

In the present paper, we demonstrate the factor analytic approach for testing the level of measurement by investigating the level of measurement of two scales based on rating

scale items, using empirical data (Study 1) and investigate the functioning of our approach using simulated data (Study 2).

Study 1: Empirical data

In this section, we present two empirical examples how to test the level of measurement. We investigate the level of measurement of (1) the *mastery goal orientation scale* with a four-point response scale and (2) the *bullying perpetration scale* with a five-point response scale.

Sample and procedure

The sample comprises 1,396 students (47.1% girls) with a mean age of 11.7 years (*SD* = 0.9) who participated in the pretest of an evaluation study (Yanagida, Strohmeier & Spiel, 2016) in Austria. Data were collected in May/June 2009 through internet-based questionnaires during one regular school hour in the schools under the supervision of two trained research assistants. Participation in the data collection was based on active parental and child consent. There were no missing data in the variables used for the present study.

Measures

Mastery Goal Orientation. The scale consists of four items from Schober, Dresel, and Ziegler (2001): (1) "I want to learn many new things", (2) "I want to do difficult things in order to learn new things", (3) "I want to understand what I am learning", (4) "I want to be able to do more and more". Cronbach's α coefficient for the mastery goal orientation scale was .82.

Answers to all questions were given on a four-point response scale ranging from 0 (*not at all*), 1 (*a little true*), 2 (*mostly true*), to 3 (*very true*). Thus, the measurement model for ordered-categorical indicators had three thresholds for each item.

Bullying Perpetration. Self-reported bullying was measured by items developed for the PISA 2009 study in Austria (Strohmeier, Gradinger, Schabmann, & Spiel, 2012). The bullying perpetration scale consists of a global item, and three specific items covering different forms of bullying. In (1) the global item, students were asked "How often have you insulted or hurt other students during the last two months?". The three specific items were similar to the global ones, except that they described specific forms of bullying, i.e., (2) verbal, (3) exclusion and (4) physical. Cronbach's α coefficient for the bullying perpetration scale was .77.

Answers to all questions were given on a five-point response scale ranging from 0 (*not at all*), 1 (*once or twice*), 2 (*two or three times a month*), 3 (*once a week*), to 4 (*nearly every day*). Thus, the measurement model for ordered-categorical indicators had four thresholds for each item of the bullying perpetration scale.

Analytic strategy

The analytic strategy follows the two steps described above, i.e., (1) estimate measurement models and (2) decide between competing models.

All models were estimated in Mplus 7.4 (Muthén & Muthén, 1998-2015) using maximum likelihood estimation with robust standard errors. For all analyses, 10 random starting values were requested for the initial and final stage optimization of the estimation procedure.

Mplus syntax for testing the level of measurement and R syntax for graphical inspection of thresholds parameters are provided in the appendices A, B, C, and D.

Results

Response category proportions for all scale items are reported in Table 1.

Step 1: Estimate measurement models. Models of the mastery goal orientation scale and bullying perpetration scale based on threshold parameter constraints for the *interval*, *rating*, and *ordinal scale model* converged properly (see Table 2).

Step 2: Decide between competing models. As for the mastery goal orientation scale, the S-B scaled chi-square difference test for the comparison *interval scale model* vs. *ordinal scale model* was statistically not significant ($\Delta\chi^2 = 12.81$, $\Delta df = 7$, $p = .077$). Moreover, the interval scale model had a lower AIC and BIC than the ordinal scale model. In addition, graphical inspection of the location of threshold parameters indicated that thresholds across all items are roughly equally spaced (see Figure 1, Panel A). In sum,

Table 1:
Response Category Proportions for Mastery Goal Orientation and Bullying Perpetration

Mastery Goal Orientation				
Response category	Item 1	Item 2	Item 3	Item 4
0 (*not at all true*)	.077	.128	.057	.090
1 (*a little true*)	.145	.203	.097	.143
2 (*mostly true*)	.223	.241	.178	.226
3 (*very true*)	.555	.428	.668	.541
Bullying Perpetration				
Response category	Item 1	Item 2	Item 3	Item 4
0 (*not at all*)	.493	.554	.768	.792
1 (*once or twice*)	.382	.342	.190	.162
2 (*two or three times a month*)	.069	.051	.021	.017
3 (*once a week*)	.027	.025	.011	.013
4 (*nearly every day*)	.029	.028	.010	.016

Table 2:
Mastery Goal Orientation and Bullying Perpetration Scale: Model Comparison between Interval Scale, Rating Scale and Ordinal Scale Model

	Mastery Goal Orientation Scale			Bullying Perpetration Scale		
Model	Interval	Rating	Ordinal	Interval	Rating	Ordinal
Number of Parameters	9	10	16	9	11	20
Loglikelihood	-5499.76	-5497.82	-5493.39	-4581.77	-4388.23	-4340.26
Scaling Correction Factor	1.04	1.07	1.02	1.62	1.26	1.04
AIC	11017.52	11015.64	11018.79	9181.54	8798.47	8720.53
BIC	11064.70	11068.06	11102.65	9228.71	8856.12	8825.36
S-B scaled χ^2 Difference Test	Interval vs. Ordinal Scale Model $\Delta\chi^2 = 12.81$, $\Delta df = 7$, $p = .077$			Interval vs. Ordinal Scale Model $\Delta\chi^2 = 854.22.19$, $\Delta df = 11$, $p < .001$		
	Rating vs. Ordinal Scale Model $\Delta\chi^2 = 9.46$, $\Delta df = 6$, $p = .149$			Rating vs. Ordinal Scale Model $\Delta\chi^2 = 124.42$, $\Delta df = 9$, $p < .001$		

Note. S-B scaled χ^2 Difference Test = Satorra-Bentler scaled chi-square difference test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

results show that the scale items of the mastery goal orientation scale are metric. The model fit of the *interval scale model* based on the underlying response variable approach was very good ($\chi^2(9) = 34.70$, $p < .001$, CFI = .994 and RMSEA = 0.045).

Since model comparison was in favor of the *interval scale model*, an additional model comparison between *rating scale model* vs. *ordinal scale model* is not required. Nevertheless, Table 2 summarizes the results of all model comparisons.

As for the bullying perpetration scale, the S-B scaled chi-square difference test for the comparison *interval scale mode*l vs. *ordinal scale model* was statistically significant ($\Delta\chi^2 = 854.22$, $\Delta df = 11$, $p < .001$). In addition, the model comparison *rating scale model* vs. *ordinal scale model* was also statistically significant ($\Delta\chi^2 = 124.42$, $\Delta df = 9$, $p < .001$). Thus, according to the chi-square difference test the *ordinal scale model* is more favorable than the *interval scale model* and the *rating scale model*. Moreover, the *ordinal scale model* had the lowest AIC and BIC. Table 2 summarizes the results of the model comparisons. In addition, graphical inspection of the location of threshold parameters indicated that thresholds across all items do not conform with the *interval scale* or the *rating scale model* (see Figure 1, Panel B). The model fit of the *ordinal scale model* based on the underlying response variable approach was very good ($\chi^2(2) = 3.34$, $p = .189$, CFI = 1.000 and RMSEA = 0.022). In sum, results show that the scale items of the bullying perpetration scale are not metric, but ordered-categorical.

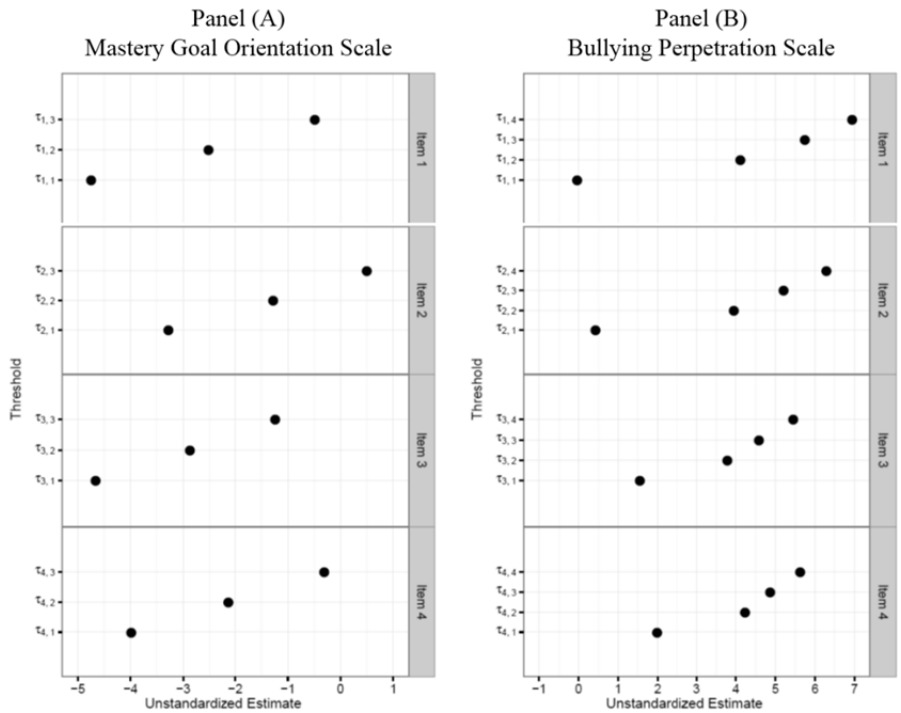

Figure 1:
Graphical representation of threshold location based on the ordinal scale model for the mastery goal orientation scale (Panel A) and bullying perpetration scale (Panel B).

Conclusion

In the empirical examples, we demonstrated how to test the level of measurement of two scales based on two consecutive steps. Results based on S-B scaled chi-square difference test, information criteria (AIC and BIC), and graphical inspection consistently showed that *interval scale model* is tenable for the mastery goal orientation scale, but not for the bullying perpetration scale. Given the answer format for bullying perpetration asking about the frequency of different forms of aggressive behavior, it is not surprising that data resulting from these items are not metric. However, using our approach it is also possible to test the level of measurement to argue on empirical grounds.

Study 2: Simulated data

In this section, we present a simulation study based on the data constellation of the empirical examples (i.e., four items with four-point and five-point response scale) to investigate the functioning of our approach for testing the level of measurement under various simulation conditions.

Simulation design

In the simulation study, we investigated various conditions with different population models and sample sizes ranging from n = 100 to 1000 with an increment of 100. Overall, we investigated a four-point response scale (i.e., three thresholds) and a five-point response scale (i.e., four thresholds) with four items. Data were simulated in accordance to the *interval scale*, *rating scale* or *ordinal scale* model. All conditions investigated in the simulation study are shown in Table 3.

Table 3:
Simulation Design: Standardized Factor Loadings and Standardized Thresholds for the Population Models for Items with a Four-Point and a Five-Point Response Scale

	Four-Point Response Scale			Five-Point Response Scale		
Model Parameters	Interval	Rating	Ordinal	Interval	Rating	Ordinal
Standardized Factor Loading						
Item 1 (λ_1)	0.40	0.60	0.60	0.60	0.60	0.60
Item 2 (λ_2)	0.60	0.60	0.60	0.60	0.60	0.60
Item 3 (λ_3)	0.60	0.60	0.60	0.60	0.60	0.60
Item 4 (λ_4)	0.60	0.60	0.60	0.60	0.60	0.60
Standardized Threshold						
Item 1						
Thresh. 1 ($\tau_{1,1}$)	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
Thresh. 2 ($\tau_{1,2}$)	-0.50	-0.50	0.00	-0.60	-0.25	0.00
Thresh. 3 ($\tau_{1,3}$)	0.00	-0.25	0.50	-0.20	0.25	0.50
Thresh. 4 ($\tau_{1,4}$)				0.20	0.50	0.75
Item 2						
Thresh. 1 ($\tau_{2,1}$)	-0.75	-0.75	-0.75	-0.80	-0.80	-0.80
Thresh. 2 ($\tau_{2,2}$)	-0.25	-0.25	0.05	-0.40	-0.05	0.00
Thresh. 3 ($\tau_{2,3}$)	0.25	0.00	0.25	0.00	0.45	0.20
Thresh. 4 ($\tau_{2,4}$)				0.40	0.70	0.60
Item 3						
Thresh. 1 ($\tau_{3,1}$)	-0.50	-0.50	-0.50	-0.40	-0.70	-0.40
Thresh. 2 ($\tau_{3,2}$)	0.00	0.00	-0.20	0.00	0.05	-0.10
Thresh. 3 ($\tau_{3,3}$)	0.50	0.25	0.40	0.40	0.55	0.50
Thresh. 4 ($\tau_{3,4}$)				0.80	0.80	0.70
Item 4						
Thresh. 1 ($\tau_{4,1}$)	0.00	0.25	-0.50	-0.20	-0.50	-0.75
Thresh. 2 ($\tau_{4,2}$)	0.50	0.75	-0.20	0.20	0.25	-0.50
Thresh. 3 ($\tau_{4,3}$)	1.00	1.00	0.40	0.60	0.75	0.00
Thresh. 4 ($\tau_{4,4}$)				1.00	1.00	1.00

In sum, 10,000 replications were conducted for each of the 2 (four point and five point scale) x 3 (interval, rating and ordinal scale model) x 10 (sample size) = 60 simulation conditions. Each replication was analyzed according to the two consecutive steps based on the factor analytic approach for testing the level of measurement. The main result of the simulation study was the model recovery rate, i.e., the percentage of the population model chosen according to the statistical test and the information criteria. In line with a significance level of $\alpha = 0.05$, a model recovery rate over 95% can be considered as an acceptable rate.

Data were simulated in R version 3.2.5 (R Development Core Team, 2016) using the lavaan package version 0.5-22 (Rosseel, 2012). All models were estimated in Mplus 7.4 (Muthén & Muthén, 1998-2015) using maximum likelihood estimation with robust standard errors based on 10 random starting values for the initial and final stage optimization.

Results

Model Non-Convergence. In all simulation conditions for the four-point and five-point response scale, there were no convergence problems, i.e., all models converged properly.

Four-point response scale. Results for the simulation study for the four-point response scale are shown in Table 4.

Table 4:
Simulation Results for the Four-Point Response Scale: Percentage an Interval Scale, Rating Scale or Ordinal Scale was Chosen Depending on the Sample Size, Statistical Test, Information Criterion and Population Model

	Population Model								
	Interval Scale Model			Rating Scale Model			Ordinal Scale Model		
Sample size (*n*)	Interval	Rating	Ordinal	Interval	Rating	Ordinal	Interval	Rating	Ordinal
n = 100									
χ^2 Difference Test	93.40%	1.46%	5.14%	24.46%	65.20%	10.34%	0.00%	0.00%	100%
AIC	81.12%	14.22%	4.66%	1.46%	91.13%	7.41%	0.00%	0.00%	100%
BIC	96.72%	3.28%	0.00%	7.99%	92.00%	0.01%	3.14%	0.19%	96.67%
n = 200									
χ^2 Difference Test	94.36%	1.20%	4.44%	2.30%	88.59%	9.11%	0.00%	0.00%	100%
AIC	80.61%	14.91%	4.48%	0.00%	91.37%	8.63%	0.00%	0.00%	100%
BIC	97.95%	2.05%	0.00%	0.26%	99.74%	0.00%	0.00%	0.00%	100%

n = 300									
χ^2 Difference Test	94.52%	1.26%	4.22%	0.13%	91.47%	8.40%	0.00%	0.00%	100%
AIC	80.95%	14.76%	4.29%	0.01%	90.87%	9.12%	0.00%	0.00%	100%
BIC	98.44%	1.56%	0.00%	0.01%	99.99%	0.00%	0.00%	0.00%	100%
n = 400									
χ^2 Difference Test	94.55%	1.27%	4.18%	0.01%	90.30%	9.69%	0.00%	0.00%	100%
AIC	81.28%	14.44%	4.28%	0.01%	89.19%	10.80%	0.00%	0.00%	100%
BIC	98.56%	1.44%	0.00%	0.01%	99.99%	0.00%	0.00%	0.00%	100%
n = 500									
χ^2 Difference Test	94.23%	1.33%	4.44%	0.00%	89.09%	10.91%	0.00%	0.00%	100%
AIC	81.7%	14.15%	4.68%	0.00%	87.52%	12.48%	0.00%	0.00%	100%
BIC	98.93%	1.07%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 600									
χ^2 Difference Test	93.75%	1.70%	4.55%	0.00%	89.27%	10.73%	0.00%	0.00%	100%
AIC	80.50%	14.78%	4.72%	0.00%	87.73%	12.27%	0.00%	0.00%	100%
BIC	98.74%	1.26%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 700									
χ^2 Difference Test	94.72%	1.16%	4.12%	0.00%	88.27%	11.73%	0.00%	0.00%	100%
AIC	81.17%	14.44%	4.39%	0.00%	86.29%	13.71%	0.00%	0.00%	100%
BIC	98.88%	1.12%	0.00%	0.00%	99.99%	0.01%	0.00%	0.00%	100%
n = 800									
χ^2 Difference Test	94.46%	1.24%	4.30%	0.00%	86.70%	13.30%	0.00%	0.00%	100%
AIC	80.38%	15.17%	4.45%	0.00%	85.10%	14.90%	0.00%	0.00%	100%
BIC	99.11%	0.89%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 900									
χ^2 Difference Test	93.79%	1.43%	4.78%	0.00%	85.99%	14.01%	0.00%	0.00%	100%
AIC	80.40%	14.62%	4.98%	0.00%	84.00%	16.00%	0.00%	0.00%	100%
BIC	99.06%	0.94%	0.00%	0.00%	100%	0%	0.00%	0.00%	100%
n = 1000									
χ^2 Difference Test	94.16%	1.47%	4.37%	0.00%	85.99%	14.01%	0.00%	0.00%	100%
AIC	80.33%	15.09%	4.58%	0.00%	84.00%	16.00%	0.00%	0.00%	100%
BIC	98.94%	1.06%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%

Note. χ^2 Difference Test = Satorra-Bentler scaled chi-square difference test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Interval scale model condition. In the *interval scale model* condition, the S-B scaled chi-square difference test and the AIC did not reach a model recovery rate of 95%, irrespective of the sample size. The BIC, on the other hand, had a model recovery rate between 96.72% (n = 100) and 99.11% (n = 800), depending on the sample size.

Rating scale model condition. In the *rating scale model* condition, the S-B scaled chi-square difference test and the AIC did not reach a model recovery rate of 95%, irrespective of the sample size. The BIC, on the other hand, had a model recovery rate over 99% when sample size was equal or larger than n = 200.

Ordinal scale model condition. In the *ordinal scale model* condition, the S-B scaled chi-square difference test and the AIC and BIC always decides for the correct model, irrespective of the sample size. The BIC, on the other hand, had a model recovery rate of 96.67% when sample size was n = 100, but always decides for the correct model when sample size was equal or larger than n = 200.

Five-point response scale. Results for the simulation study for the five-point response scale show a similar pattern like the results of the four-point response scale and are shown in Table 5.

Table 5:
Simulation Results for the Five-Point Response Scale: Percentage an Interval Scale, Rating Scale or Ordinal Scale was Chosen Depending on the Sample Size, Statistical Test, Information Criterion and Population Model

	Population Model								
	Interval Scale Model			Rating Scale Model			Ordinal Scale Model		
Sample size (n)	Interval	Rating	Ordinal	Interval	Rating	Ordinal	Interval	Rating	Ordinal
n = 100									
χ^2 Difference Test	93.74%	1.66%	4.60%	0.43%	89.79%	9.76%	0.06%	0.01%	99.93%
AIC	84.83%	12.93%	2.24%	0.00%	94.35%	5.65%	0.04%	0.01%	99.95%
BIC	98.95%	1.05%	0.00%	0.06%	99.94%	0.00%	8.64%	0.67%	90.69%
n = 200									
χ^2 Difference Test	94.28%	1.71%	4.01%	0.00%	89.62%	10.38%	0.00%	0.00%	100%
AIC	84.72%	13.30%	1.98%	0.00%	93.00%	7.00%	0.00%	0.00%	100%
BIC	99.53%	0.47%	0.00%	0.00%	100%	0.00%	0.03%	0.00%	99.97%
n = 300									
χ^2 Difference Test	94.09%	1.74%	4.17%	0.01%	87.89%	12.10%	0.00%	0.00%	100%
AIC	84.27%	13.63%	2.10%	0.00%	91.15%	8.85%	0.00%	0.00%	100%
BIC	99.54%	0.46%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%

n = 400									
χ^2 Difference Test	94.36%	1.76%	3.88%	0.00%	85.39%	14.61%	0.00%	0.00%	100%
AIC	84.93%	13.06%	2.01%	0.00%	88.98%	11.02%	0.00%	0.00%	100%
BIC	99.76%	0.24%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 500									
χ^2 Difference Test	93.96%	1.93%	4.11%	0.00%	81.11%	18.89%	0.00%	0.00%	100%
AIC	84.27%	13.58%	2.15%	0.00%	85.77%	14.23%	0.00%	0.00%	100%
BIC	99.83%	0.17%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 600									
χ^2 Difference Test	94.36%	1.79%	3.85%	0.00%	79.38%	20.62%	0.00%	0.00%	100%
AIC	84.26%	13.75%	1.99%	0.00%	84.00%	16.00%	0.00%	0.00%	100%
BIC	99.88%	0.12%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 700									
χ^2 Difference Test	93.95%	0.73%	4.32%	0.00%	75.74%	24.26%	0.00%	0.00%	100%
AIC	84.25%	13.64%	2.11%	0.00%	80.71%	19.29%	0.00%	0.00%	100%
BIC	99.81%	0.19%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 800									
χ^2 Difference Test	94.47%	1.46%	4.07%	0.00%	71.91%	28.09%	0.00%	0.00%	100%
AIC	85.51%	12.52%	1.97%	0.00%	77.40%	22.60%	0.00%	0.00%	100%
BIC	99.89%	0.11%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 900									
χ^2 Difference Test	93.88%	1.87%	4.25%	0.00%	69.41%	30.59%	0.00%	0.00%	100%
AIC	84.27%	13.53%	2.20%	0.00%	75.27%	24.73%	0.00%	0.00%	100%
BIC	99.82%	0.18%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%
n = 1000									
χ^2 Difference Test	94.11%	1.84%	4.05%	0.00%	66.44%	33.56%	0.00%	0.00%	100%
AIC	83.98%	13.92%	2.10%	0.00%	72.58%	27.42%	0.00%	0.00%	100%
BIC	99.93%	0.07%	0.00%	0.00%	100%	0.00%	0.00%	0.00%	100%

Note. χ^2 Difference Test = Satorra-Bentler scaled chi-square difference test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

In summary, S-B scaled chi-square difference test and the AIC did not reach a model recovery rate of 95% in the *interval* and *rating scale model* irrespective of the sample size, whereas the BIC always had a model recovery rate over 95% in the *interval* and *rating scale model*. As for the *ordinal scale model*, the S-B scaled chi-square difference test and the AIC nearly always decides for the correct model irrespective of the sample size. The BIC, on the other hand, had a model recovery rate of 90.69% when sample size

was $n = 100$, but nearly always decided for the correct model when sample size was equal or larger than $n = 200$.

Conclusion

Results of the simulation study demonstrate that the factor analytic approach for testing the level of measurement is functioning in principle. Moreover, results show that the BIC is the best criterion to decide between competing models. This criterion had a model recovery rate of over 95% in all simulation conditions as long as sample size was equal or larger than $n = 200$. The S-B scaled chi-square difference test and the AIC, however, never reached a model recovery rate of 95% in the *interval* and *rating scale model* irrespective of the sample size.

Discussion

The present study presented a factor analytic approach for testing the level of measurement of variables stemming from scale items with various answer formats. This approach is based on the *response function approach* for measurement models for ordered-categorical indicators and enables to statistical compare the *interval scale* and the *ordinal scale model*. The decision for the *interval scale model* indicates that data are consistent with the assumption of equidistance of response categories (i.e., variables are metric). On the other hand, the decision for the *ordinal scale model* indicates that data stemming from scale items are not interval but ordinal level. In case the comparison is in favor of the *ordinal scale model*, an additional comparison between the *rating scale* and the *ordinal scale model* is recommended since the *rating scale model* is also accounting for the ordinal nature of scale items, but is a more parsimonious model than the *ordinal scale model.* Thus, in structural equation modeling a *rating scale model* can be specified for the measurement part of the model to save degrees of freedom.

In order to decide between competing models, we propose using the Bayesian information criterion (BIC) and a graphically inspection of threshold parameters whether thresholds spacing across all items are in line with the *interval scale*, *rating scale* or *ordinal scale model*. Results of the simulation study showed that the BIC has a model recovery rate of over 95% in all simulation conditions as long as sample size was equal or larger than $n = 200$. Of course, these recommendations are solely based on the simulation conditions investigated in the present study. Additional simulation studies based on different population models, i.e., with varying the factor loadings, the number of thresholds, the number of items, and the degree of model deviation and the sample size, should be conducted to further investigate the factor analytic approach for testing the level of measurement. Moreover, the present study compared three basic models (*interval scale*, *rating scale* and *ordinal scale model*), whereas further models with thresholds spacing based on other hypotheses might be investigated.

In conclusion, we recommend testing the level of measurement using the factor analytic approach instead of simply assuming variables stemming from scale items are metric

when applying statistical methods, which in fact require metric data. In such a way, biased parameter estimates and wrong research conclusions can be avoided, as ignoring the level of measurement is known to be a measurement error fallacy that is most consequential and prevalent in published quantitative research (see Wang et al., 2013). Of course, in case where the answer format clearly suggests an ordered-categorical (e.g., less than five response options) or metric level of measurement (e.g., visual analog scale), statistical methods for testing this assumption are not needed. Note that a simulation study showed that the performance of robust continuous (e.g., normal theory maximum likelihood with robust correction) and categorical methodology (e.g., robust categorical least squares) in estimating confirmatory factor analysis models depends on the number of categories (Rhemtulla, Brosseau-Liard & Savalei, 2012). That is, when observed variables have fewer than five response categories, robust categorical methodology is best, whereas both methods yield acceptable performance with five to seven response categories. These results indicate that variables stemming from scale items may be treated as continuous as long as they have at least five response categories.

Of course, there are other statistical assumptions (e.g., distributional assumptions) which also need to be considered, but were not in the scope of the present paper. Thus, beside the level of measurement it is important to always investigate and tackle all assumptions when applying statistical methods.

Competing interests

There are no competing interests.

Funding

The data analyses and writing of the present study was funded by the Platform for Intercultural Competences, University of Applied Sciences Upper Austria (PI: Dagmar Strohmeier).

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105-113.

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495-511). New York, NY: Guilford Press.

Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195-208). New York, NY: Guilford Press.

Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests. [Introduction into theory of psychological tests]. Bern: Huber.

Hohensinn, C., & Kubinger, K. D. (2017). Using Rasch model generalizations for taking testees' speed, in addition to their power, into account. *Psychological Test and Assessment Modeling*, *59*, 93-108.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*, 347-387. doi: 10.1207/S1532790 6347-387

Li., P-W., & Wu., Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495-511). New York, NY: Guilford Press.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Erlbaum.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*, 48-65.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide*. Seventh edition. Los Angeles, CA: Muthén & Muthén.

Pett, M. A. (2015). *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions*. Thousand Oaks, CA: Sage Publications.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation method under a range of non-ideal situations. *Psychological Methods*, *17*, 354-373.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1-36. doi: 10.18637/jss.v048.i02

Schober, B., Dresel, M., & Ziegler, A. (2001). *Skalen zur Erfassung der Motivationalen Orientierung im Fach Mathematik* [Scales for measuring motivational orientation in mathematics]. Unpublished manuscript, Ludwig-Maximilians-Universität München, Munich, Germany.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A., & Rabe-Hesketh, S. (2005). Structural equation modeling: Categorical variables. In B. Everitt, & D. C. Howell (Ed.), *Encyclopedia of statistics in behavioral science.* London: Wiley.

Strohmeier, D., Gradinger, P., Schabmann, A., & Spiel, C. (2012). Gewalterfahrungen von Jugendlichen. Prävalenzen und Risikogruppen [Adolescent's experiences of violence]. In F. Eder (Ed.) *PISA 2009. Nationale Zusatzerhebungen* [PISA 2009. Additional National Investigations] (pp. 165–208). Münster: Waxmann.

Stevens, S. S. (1960). On the theory of scales of measurement. In A. Danto & S. Morgenbesser (Eds.), *Philosophy of science* (pp. 141-149). New York: Meridian.

Wang, L., Watts, A. S., Anderson, R. A., & Little, T. D. (2013). Common fallacies in quantitative research methodology. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods, Volume 2: Statistical analysis* (pp. 718-758). New York, NY: Oxford University Press.

West, S. G., Taylor A. B., Wu W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.). *Handbook of structural equation modeling* (pp. 209-231). New York: Guildford Press.

Yanagida, T., Strohmeier, D., & Spiel, C. (in press). Dynamic Change of Aggressive Behavior and Victimization among Adolescents: Effectiveness of the ViSC Program. *Clinical Child and Adolescent Psychology*.

Appendix A: Mplus Syntax for the Four-Point Response Scale

Syntax for interval scale

```
TITLE:      Four-Point Response Scale
            Interval Scale Model

DATA:       FILE IS Data.dat;

VARIABLE:   NAMES ARE item1 item2 item3 item4;
            CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:   ESTIMATOR = MLR;
            STARTS = 10;

MODEL:      f BY item1* item2 item3 item4;

            f@1;

            [item1$1] (T11);
            [item1$2] (T12);
            [item1$3] (T13);

            [item2$1] (T21);
            [item2$2] (T22);
            [item2$3] (T23);

            [item3$1] (T31);
            [item3$2] (T32);
```

```
              [item3$3] (T33);

              [item4$1] (T41);
              [item4$2] (T42);
              [item4$3] (T43);

MODEL CONSTRAINT:
              NEW(diff);

              diff = T12 - T11;

              T12 = T13 - diff;

              T21 = T22 - diff;
              T22 = T23 - diff;

              T31 = T32 - diff;
              T32 = T33 - diff;

              T41 = T42 - diff;
              T42 = T43 - diff;

OUTPUT:       NOCHISQUARE;
```

Syntax for the rating scale model

```
TITLE:        Four-Point Response Scale
              Rating Scale Model

DATA:         FILE IS Data.dat;

VARIABLE:     NAMES ARE item1 item2 item3 item4;
              CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:     ESTIMATOR = MLR;
              STARTS = 10;

MODEL:        f BY item1* item2 item3 item4;

              f@1;

              [item1$1] (T11);
              [item1$2] (T12);
              [item1$3] (T13);

              [item2$1] (T21);
              [item2$2] (T22);
```

```
              [item2$3] (T23);

              [item3$1] (T31);
              [item3$2] (T32);
              [item3$3] (T33);

              [item4$1] (T41);
              [item4$2] (T42);
              [item4$3] (T43);

MODEL CONSTRAINT:
              NEW(diff1 diff2);

              diff1 = T12 - T11;
              diff2 = T13 - T12;

              T21 = T22 - diff1;
              T22 = T23 - diff2;

              T31 = T32 - diff1;
              T32 = T33 - diff2;

              T41 = T42 - diff1;
              T42 = T43 - diff2;

OUTPUT:       NOCHISQUARE;
```

Syntax for the ordinal scale model

```
TITLE:        Four-Point Response Scale
              Ordinal Scale Model

DATA:         FILE IS Data.dat;

VARIABLE:     NAMES ARE item1 item2 item3 item4;
              CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:     ESTIMATOR = MLR;
              STARTS = 10;

MODEL:        f BY item1* item2 item3 item4;

              f@1;

OUTPUT:       NOCHISQUARE;
```

Appendix B: Mplus Syntax for the Five-Point Response Scale

Syntax for interval scale

```
TITLE:        Five-Point Response Scale
              Interval Scale Model

DATA:         FILE IS Data.dat;

VARIABLE:     NAMES ARE item1 item2 item3 item4;
              CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:     ESTIMATOR = MLR;
              STARTS = 10;

MODEL:        f BY item1* item2 item3 item4;

              f@1;

              [item1$1] (T11);
              [item1$2] (T12);
              [item1$3] (T13);
              [item1$4] (T14);

              [item2$1] (T21);
              [item2$2] (T22);
              [item2$3] (T23);
              [item2$4] (T24);

              [item3$1] (T31);
              [item3$2] (T32);
              [item3$3] (T33);
              [item3$4] (T34);

              [item4$1] (T41);
              [item4$2] (T42);
              [item4$3] (T43);
              [item4$4] (T44);

MODEL CONSTRAINT:
              NEW(diff);

              diff = T12 - T11;

              T12 = T13 - diff;
              T13 = T14 - diff;
```

```
              T21 = T22 - diff;
              T22 = T23 - diff;
              T23 = T24 - diff;

              T31 = T32 - diff;
              T32 = T33 - diff;
              T33 = T34 - diff;

              T41 = T42 - diff;
              T42 = T43 - diff;
              T43 = T44 - diff;

OUTPUT:       NOCHISQUARE;
```

Syntax for the rating scale model

```
TITLE:        Five-Point Response Scale
              Rating Scale Model

DATA:         FILE IS Data.dat;

VARIABLE:     NAMES ARE item1 item2 item3 item4;
              CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:     ESTIMATOR = MLR;
              STARTS = 10;

MODEL:        f BY item1* item2 item3 item4;

              f@1;

              [item1$1] (T11);
              [item1$2] (T12);
              [item1$3] (T13);
              [item1$4] (T14);

              [item2$1] (T21);
              [item2$2] (T22);
              [item2$3] (T23);
              [item2$4] (T24);

              [item3$1] (T31);
              [item3$2] (T32);
              [item3$3] (T33);
              [item3$4] (T34);

              [item4$1] (T41);
```

```
              [item4$2] (T42);
              [item4$3] (T43);
              [item4$4] (T44);

MODEL CONSTRAINT:
              NEW(diff1 diff2 diff3);

              diff1 = T12 - T11;
              diff2 = T13 - T12;
              diff3 = T14 - T13;

              T21 = T22 - diff1;
              T22 = T23 - diff2;
              T23 = T24 - diff3;

              T31 = T32 - diff1;
              T32 = T33 - diff2;
              T33 = T34 - diff3;

              T41 = T42 - diff1;
              T42 = T43 - diff2;
              T43 = T44 - diff3;

OUTPUT:       NOCHISQUARE;
```

Syntax for the ordinal scale model

```
TITLE:        Five-Point Response Scale
              Ordinal Scale Model

DATA:         FILE IS Data.dat;

VARIABLE:     NAMES ARE item1 item2 item3 item4;
              CATEGORICAL ARE item1 item2 item3 item4;

ANALYSIS:     ESTIMATOR = MLR;
              STARTS = 10;

MODEL:        f BY item1* item2 item3 item4;

              f@1;

OUTPUT:       NOCHISQUARE;
```

Appendix C: R Syntax for Graphical Inspection of Thresholds Parameters of the Four-Point Response Scale

```
# install packages
install.packages(c("MplusAutomation", "ggplot2"))

# load packages
library(MplusAutomation)
library(ggplot2)

# set the working directory
# contains only the Mplus output file of the ordinal scale
model
setwd("C:/...")

# extract unstandardized model parameters
modpar <- extractModelParameters()$unstandardized

# extract thresholds
modpar <- modpar[modpar$paramHeader == "Thresholds", ]

# create data frame
df <- data.frame(item = factor(rep(paste("Item", 1:4), each
= 3)),
                 thres = factor(rep(1:3, times = 4)),
                 est = modpar$est)

# plot thresholds
ggplot(df, aes(est, thres)) + geom_point() + fac-
et_grid(item ~ .) +
   scale_y_discrete("Threshold") + xlab("Unstandardized
Estimate")
```

Appendix D: R Syntax for Graphical Inspection of Thresholds Parameters of the Five-Point Response Scale

```
# install packages
install.packages(c("MplusAutomation", "ggplot2"))

# load packages
library(MplusAutomation)
library(ggplot2)

# set the working directory
# contains only the Mplus output file of the ordinal scale
model
setwd("C:/...")

# extract unstandardized model parameters
modpar <- extractModelParameters()$unstandardized

# extract thresholds
modpar <- modpar[modpar$paramHeader == "Thresholds", ]

# create data frame
df <- data.frame(item = factor(rep(paste("Item", 1:4), each
= 4)),
                 thres = factor(rep(1:4, times = 4)),
                 est = modpar$est)

# plot thresholds
ggplot(df, aes(est, thres)) + geom_point() + fac-
et_grid(item ~ .) +
   scale_y_discrete("Threshold") + xlab("Unstandardized
Estimate")
```