Article

Applied Psychological Measurement
2016, Vol. 40(2) 128–141
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615611164
apm.sagepub.com
SAGE

Identifying Person-Fit Latent Classes, and Explanation of Categorical and Continuous Person Misfit

Judith M. Conijn[1], **Klaas Sijtsma**[2], **and Wilco H. M. Emons**[2]

Abstract

Latent class (LC) cluster analysis of a set of subscale l_z person-fit statistics was proposed to explain person misfit on multiscale measures. The proposed explanatory LC person-fit analysis was used to analyze data of students (N = 91,648) on the nine-subscale School Attitude Questionnaire Internet (SAQI). Inspection of the class-specific l_z mean and variance structure combined with explanatory analysis of class membership showed that the data included a poor-fit class, a class showing good fit combined with social desirability bias, a good-fit class, and two classes that were more difficult to interpret. A comparison of multinomial logistic regression predicting class membership and multiple regression predicting continuous person fit showed that LC cluster analysis provided information about aberrant responding unattainable by means of linear multiple regression. It was concluded that LC person-fit analysis has added value to common approaches to explaining aberrant responding to multiscale measures.

Keywords

aberrant responding, explanatory person-fit analysis, latent class cluster analysis, person-fit statistic l_z, response inconsistency, School Attitude Questionnaire Internet

Person-fit analysis (PFA; e.g., Meijer & Sijtsma, 2001) is used to investigate inconsistencies in patterns of a person's item scores on a test or a questionnaire. PFA can be confirmatory or explanatory. Confirmatory PFA aims at detecting individuals whose item-score pattern is significantly different from what is expected from the hypothesized item response model. It warns test users, such as the teacher and the clinician, that the validity of the test score may be questionable. Follow-up analysis of misfitting item-score patterns may diagnose the cause of misfit (e.g., Emons, Sijtsma, & Meijer, 2005; Ferrando, 2014). Explanatory PFA focuses on explaining inter-individual differences in response consistency. Results from explanatory PFA are useful to evaluate the validity and the feasibility of a test or a questionnaire in subgroups. For example, Conijn, Emons, De Jong, and Sijtsma (2015) showed that a generic mental health

[1] Leiden University, The Netherlands
[2] Tilburg University, The Netherlands

Corresponding Author:
Judith M. Conijn, Department of Clinical Psychology, Institute of Psychology, Leiden University, Pieter de la Court gebouw, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands.
Email: j.m.conijn@fsw.leidenuniv.nl

questionnaire may be less suitable for patients exhibiting qualitatively more distinct symptoms than patients showing common symptoms of anxiety and depression.

The literature discusses approximately 40 different person-fit statistics (Karabatsos, 2003; Meijer & Sijtsma, 2001). The most popular person-fit statistic is l_z (Drasgow, Levine, & McLaughlin, 1987). Compared with other person-fit statistics, statistic l_z has a better detection rate, especially for ordered categorical data (Emons, 2008), and hence is appropriate to flag misfitting individual item-score patterns in confirmatory PFA. In an explanatory PFA, linear regression (LR) may be used to relate continuous statistic l_z to explanatory variables for person misfit (e.g., Conijn, Emons, Van Assen, Pedersen, & Sijtsma, 2013; Reise & Waller, 1993).

Statistic l_z is suitable for unidimensional scales and has more power as scale length increases. This implies that when a questionnaire consists of several short scales, PFA using l_z has to focus on each separate, short scale and thus has to deal with limited amounts of information, thus rendering l_z an unreliable and powerless PFA measure (Emons, 2008). This unfortunate situation can be avoided by combining person-fit information from different, short scales into a single person-fit statistic by taking the sum of the separate l_z values, resulting in statistic l_{zm} (Drasgow, Levine, & McLaughlin, 1991) which summarizes overall fit. Statistic l_{zm} may be tested for significance or regressed on a set of explanatory variables, thus increasing the precision of the confirmatory or explanatory PFA (Conijn, Emons, & Sijtsma, 2014). However, for both confirmatory and explanatory purposes, comprehensive statistic l_{zm} may not work well if misfit is present only for some subscales but not for others. Conijn et al. (2014) proposed alternatives for confirmatory PFA on multiple-scales tests by combining subscale-level person information.

This study focuses on explanatory PFA for tests with multiple subscales. Using an example from education, the authors introduce the latent class (LC) cluster analysis approach to PFA (LC-PFA), which quantifies person fit by assigning the individuals to one of K nominal latent person-fit classes. For example, one latent person-fit class is found when all students show good fit to all subscales, two person-fit classes when one student subgroup fits well and another subgroup fits poorly, and multiple person-fit classes when several subgroups show different misfit patterns. Person-fit classes can reveal both quantitative and qualitative inter-individual differences in response consistency. In contrast, the comprehensive l_{zm} statistic quantifies person fit by assigning individuals a person-fit value on a continuous unidimensional scale reflecting only differences in degree, and may fail to detect subscale-specific misfit. LC-PFA thus may alleviate the problem of unreliable l_z values based on short subscales, and provide a finer-grained description of person fit across multiple scales than the comprehensive l_{zm} statistic. The usefulness of LC-PFA was explored in an application to data of the nine subscales of the School Attitude Questionnaire Internet (SAQI) (Vorst, Smits, Oort, Stouthard, & David, 2008).

LC-PFA

For each SAQI subscale, statistic l_z^p for polytomous items (Drasgow, Levine, & Williams, 1985) was used to quantify person fit relative to the graded response model (GRM; Samejima, 1997). A bootstrap procedure (e.g., De la Torre & Deng, 2008) was used to obtain normally distributed l_z^p values under the null model of response consistency with the GRM. Larger negative l_z^p and l_{zm}^p values suggest a higher degree of misfit; statistic l_{zm}^p is the standardized sum of subscale l_z^p values (Conijn et al., 2014; Hulin, Drasgow, & Parsons, 1983).

LC-PFA works as follows. Let $\mathbf{l}_i^p = (l_{z1i}^p, \ldots, l_{zQi}^p)$ be the vector of the Q subscale l_z^p values of person i. The LC cluster model describes the distribution of $\mathbf{l}_i^p$ as a mixture of K class-specific distributions of $\mathbf{l}_i^p$ $(i = 1, \ldots, N)$. Let ξ_k $(k = 1, \ldots, K)$ be the vector of means, variances, and the covariances of the l_z^p values in each class k, and let π_k denote the relative class size. The LC-PFA model is defined as

$$f(\mathbf{l}_i^p) = \sum_{k=1}^{K} \pi_k f(\mathbf{l}_i^p | \xi_k).$$

In an explanatory context, several LC cluster models are estimated, starting with the 1-class LC cluster model, the 2-class model, and so on, until a model is identified with relative good model fit. Students are assigned to a person-fit class using modal classification. Class membership is then regressed on explanatory variables by means of multinomial logistic regression (MLR) using Vermunt's (2010) three-step approach. This approach takes the uncertainty of modal classification into account and thus produces more accurate estimates than standard MLR of class membership.

Application of LC-PFA to the SAQI

In the Netherlands, the SAQI is an important instrument used to identify students with low social and emotional well-being at both elementary and high school. Different causes of invalid responses to the SAQI producing misfit are conceivable. Examples are insufficient concentration, motivation, self-knowledge, and language skills, which might be revealed by separate classes. For example, LC-PFA may identify a class of young children showing misfit to subscales addressing complicated ''adult'' traits, indicating low traitedness (Reise & Waller, 1993). LC-PFA may also identify person consistent misfit to all SAQI subscales due to lack of motivation.

Three research goals were addressed. First, the authors assessed the number of person-fit classes in the SAQI data. A LC solution including a good-fit class, a misfit class, and additional latent classes representing subscale-specific person misfit, was expected. Second, the authors assessed how the latent person-fit classes relate to explanatory variables of person misfit such as age, education level, SAQI completion time, concentration in class, and socially desirable responding. Third, using the same explanatory variables, the explanatory PFA results from MLR of class membership were compared with the results from LR of continuous multiscale statistic l_{zm}^p. Results showing (a) a well-interpretable LC cluster solution, (b) meaningful relationships between LCs and explanatory variables for person misfit, and (c) additional insights into aberrant responding to the SAQI compared with those found by means of LR, suggest that LC-PFA is a valuable approach to explanatory PFA.

Method

Participants

Permission was obtained for secondary analyses of data from 91,648 participants (50.6% males) who completed the SAQI online at 528 different elementary schools and high schools in the Netherlands. Age ranged from 8 to 16 years (M = 12.48; SD = 1.19), and 15.7% of the students attended elementary school (age range = 8-13 years). At Dutch high schools, the increasing levels of education are VMBO (i.e., preparatory middle-level vocational education), HAVO (i.e., general secondary education), and VWO (i.e., pre-university secondary education). The percentage of students was 39.7% for VMBO, 10.0% for combined VMBO/HAVO, 6.3% for HAVO, 17.4% for combined HAVO/VWO, and 10.9% for VWO. Within VMBO five types of education can be distinguished, but one group, 'praktijkonderwijs' (literally 'practical education') was excluded from our sample because earlier research (Vorst et al., 2008) found their scores to be inconsistent with theoretical SAQI models.

Measures

The nine SAQI subscales together measure three educational attitudes: Motivation to perform school tasks is measured by means of the subscales Task Attitude, Concentration in Class, and Homework Attitude; satisfaction with life at school is measured by means of the subscales Fun at School, Feeling Socially Accepted (henceforth called: ‘‘Social Acceptance’’), and Relationship with Teachers; and self-confidence with one’s own school-related competencies is measured by means of the subscales Expression Skills, Exam Confidence, and Social Competence. The SAQI also includes a Social Desirability bias scale.

Each SAQI subscale has two parallel versions consisting of eight items each. Students completed both versions in a single administration. Items had 3-point rating scales that were scored 1 (*no*), 2 (*don’t know*), or 3 (*yes*). The SAQI subscales are balanced; that is, half of the items are positively worded and the other items are negatively worded. In the current sample, alpha coefficients for the SAQI subscales ranged from .80 to .90.

To interpret the latent person-fit classes, the explanatory variables age, education level, language grade, SAQI subscale total scores, and SAQI completion time were used. Explanatory variables were also included for quantifying response scale use. They were the frequency of ‘‘yes’’, ‘‘don’t know’’, and ‘‘no’’ responses (before recoding), and the frequency of ‘‘3’’ scores (after recoding), computed across the 144 items of the nine SAQI scales. Due to balanced SAQI subscales, a high frequency of ‘‘yes’’ responses combined with a low frequency of ‘‘no’’ responses suggests agreement bias, whereas the opposite pattern suggests disagreement bias. SAQI traits are all socially desirable, thus a high frequency of ‘‘3’’ scores may reflect social desirability bias.

Statistical Analyses

GRM fit of each of the SAQI subscales was assessed by testing GRM model assumptions. In a first round of model-fit evaluation, factor analysis results were used to assess whether the two parallel versions of each of the SAQI subscales could be merged to double test length and gain PFA power. It was found that merging parallel subscales did not deteriorate model fit; hence, each of the parallel subscale pairs was treated as a single subscale in all further analysis. For testing the assumptions of unidimensionality and local independence, factor analysis for categorical data in Mplus (Muthén & Muthén, 2007) was performed. Values of the root mean square error approximation (RMSEA) $\leq$ .08 and the Tucker–Lewis index (TLI) $\geq$ .95 indicate satisfactory fit. It was found that RMSEA (range = .09 - .13) suggested poor fit for the combined Concentration in the Class, Homework Attitude, Pleasure at School, Exam Confidence, and Social Competence subscales but satisfactory fit (range = .04 - .08) for the other four subscales. TLI suggested satisfactory fit for all nine subscales (range = .92 - .98). For assessing the logistic shape of the item step response functions, graphical analysis was used. No substantial deviations from monotone increasing item step response functions were found. To summarize, for some subscales, evidence of mild model misfit was found. However, because statistic l_z^p is robust against mild model misfit (Conijn et al., 2014), statistic l_z^p was computed for each pair of parallel subscales.

The resulting nine subscale l_z^p values were used to perform LC cluster analysis using Latent Gold (Vermunt & Magidson, 2005). The authors estimated the one-class LC cluster model, the two-class model, and so on, until the nine-class model. For each LC solution, three versions were estimated: (a) a model in which only the l_z^p means were allowed to differ across classes, whereas l_z^p variances were held equal across classes, and covariances between l_z^p values of different scales were fixed at 0; (b) as Model 1, but l_z^p variances were also allowed to vary across classes; and (c) as Model 2, but covariances were estimated freely while keeping them equal across classes.

To choose the best fitting model, the Bayesian information criterion (BIC), the proportion of classification errors (CEs) reflecting the expected proportion of misclassified cases based on LC cluster model parameters (i.e., class size π_k and class-specific mean and covariance structure ξ_k), and the bivariate residuals were used. The authors further inspected entropy value R^2, which is a pseudo squared multiple correlation measure quantifying how well the observed variables in the model predict the LCs (Vermunt & Magidson, 2005). Lower BIC and CE and higher R^2 suggest better model fit. Class-specific l_z^p means were inspected in order to select an LC solution having sufficiently distinct classes. Cross-validation using four random subsamples allowed assessing the stability of the LC cluster model.

Explanatory PFA

MLR was used to investigate the association between assigned class membership and explanatory variables age, gender, education level, language grade, SAQI completion time, and several SAQI subscale scores expected to be related to person misfit: Social Desirability, Concentration in the Class, Task Attitude, and Expression Skills. MLR results were compared with LR results on the same explanatory variables. The continuous explanatory variables were standardized.

In LR, explanatory variables' unique contribution to the model was determined by R^2 decrease caused by excluding the explanatory variable of interest from the model. In MLR, unique contributions were assessed using pseudo R^2 and the CE under the prediction model, which differs from the CE used for model selection. The CE under the prediction model is an estimate of the proportion of CEs when persons are assigned to the latent person-fit classes based on the explanatory variables only. Statistic CE used for model selection is an estimate of the percentage of CEs when both person-fit values l_z^p and the explanatory variables (if they are included in the model) are used to assign persons to the latent person-fit classes. Finally, it is noted that pseudo R^2 is not a proportion of explained variance and its interpretation is therefore different from the R^2 in LR.

Results

Descriptive Statistics

Except for the Exam Confidence scores, the other eight SAQI subscale scores were negatively skewed. Skewness ranged from –2.03 to –0.45. Thirty-five percent of the students interrupted test taking while registration timing was continued, rendering completion time useless which then was treated as missing. Given that the total number of items equaled 144 and the SAQI subscales are balanced, the average frequencies of ''yes'' (M = 64; SD = 12), ''no'' (M = 50; SD = 12), and ''don't know'' (M = 29; SD = 20) responses suggested a tendency to agree (i.e., agreement bias). Average 3-score frequency also was high (M = 95; SD = 27), suggesting either a general positive educational attitude or a tendency to socially desirable response bias. Average completion time was 19 min (SD = 3) and average language grade was 7.05 (SD = 1; potential range = 1-10). The correlations between l_z^p values ranged from .04 to .23, suggesting that students did not consistently show good fit or poor fit across all nine SAQI subscales. Based on significance level α = .05, multiscale person-fit statistic l_{zm}^p classified 7.0% of the item-score patterns as misfitting.

Person-Fit Classes in the SAQI Data

Model comparison. LC cluster model comparisons showed that there were large differences between l_z^p variances across classes, but that l_z^p covariances were small. Therefore, models in which l_z^p means and variances were estimated freely across classes and covariances were fixed at 0, were most appropriate for further inspection. Table 1 shows the BIC and CE values for these models.

BIC decreased with every additional class and CE was at least .15 for all models having more than two classes. For each of the models, almost all bivariate residuals were significant, suggesting poor absolute model fit. However, going from four to five classes produced a noticeable decrease of BIC and CE values. Entropy R^2 equaled .73 for the five-class model, suggesting moderately good separation between classes (Vermunt, 2010). To conclude, the five-class model provided the best trade-off between model complexity and model fit; hence, the five-class model was retained for further analysis.

Model interpretation. For the five-class model, Table 2 shows class sizes, denoted by π, and l_z^p mean and variance estimates. For each subscale, Class 1 ($\pi_1 = .10$) had a negative mean l_z^p value which was also lowest compared with the other classes. The l_z^p variance estimates of Class 1 were the highest of all classes. Class 2 ($\pi_2 = .25$) had only positive l_z^p mean estimates. The l_z^p mean for Social Acceptance was the highest across classes and the variance equaled 0 due to all students having the maximum Social Acceptance score. Both Class 3 ($\pi_3 = .20$) and Class 4 ($\pi_4 = .25$) had variable l_z^p means but overall person fit was better for Class 3 than for Class 4, particularly with respect to the Exam Confidence subscale. Class 5 ($\pi_5 = .21$) was similar to Class 2 in terms of estimated means and variances. Except for the Social Acceptance l_z^p, Class 5 had the highest l_z^p mean and the lowest l_z^p variance for each subscale across all classes.

Based on the class-specific l_z^p structure, classes were characterized as follows. Class 1: poor fit; Class 2: good fit, perfect Social Acceptance; Class 3: mixed-good fit; Class 4: mixed fit; and Class 5: good fit. The interpretation was consistent with l_{zm}^p classification based on $\alpha = .05$. Statistic l_{zm}^p was significant for 61% of the Class 1 students, 2% of the Class 3 students, 5% of Class 4 students, and 0% of the students in Classes 2 and 5.

Model cross-validation. The stability of the five-class solution was assessed by re-estimating all nine models in four subsamples ($n = 22{,}912$; see Table 1). In three subsamples, BIC and CE suggested that three, four, or seven classes were more appropriate than five classes (Table 1, BIC values bold faced), but closer inspection of those models showed that they always included a class highly similar to Class 2 (good fit, perfect Social Acceptance) and Class 5 (good fit) in the five-class model. Three of the five models that fitted relatively well also provided a class highly similar to Class 1 (poor fit), and two models provided a class similar to Class 4 (mixed fit). To conclude, cross-validation results suggested that the optimal number of classes may vary across samples, but that the results with respect to explanatory PFA remain largely intact; hence, it may be reasonable to maintain the five-class model.

Relationships Between Person-Fit Classes and Explanatory Variables

Preliminaries. First, profile plots of the class-specific mean values of the explanatory variables were inspected. Next, MLR was conducted to estimate the partial effects of explanatory variables (Table 3, Column 1) on class membership. The variables and the corresponding hypotheses were derived from previous research: Males fit worse than females (Pinsoneault, 2002); self-knowledge increases with age, and hence, age relates negatively to misfit (Meijer, Egberink, Emons, & Sijtsma, 2008); education level, completion time, concentration, and task

Table 1. Model-Fit Indices for Latent Class Cluster Models With Freely Estimated Means and Variances.

		Total sample			Subsample 1		Subsample 2		Subsample 3		Subsample 4	
No. of classes	No. of parameters	Log likelihood	BIC	CE	BIC	CE	BIC	CE	BIC	CE	BIC	CE
1	18	−995,865	1,991,936	0.00	492,613	0.00	493,273	0.00	484,557	0.00	490,341	0.00
2	37	−937,255	1,874,932	0.08	464,669	0.09	464,828	0.08	456,766	0.09	461,776	0.09
3	56	−926,997	1,854,634	0.15	**428,291**	**0.07**	460,201	0.15	452,102	0.15	456,734	0.15
4	75	−919,229	1,839,317	0.23	423,040	0.13	456,483	0.23	442,238	0.16	**420,249**	**0.12**
5	94	**−829,798**	**1,660,670**	**0.17**	419,038	0.17	454,576	0.26	**413,546**	**0.17**	416,265	0.17
6	113	−823,920	1,649,133	0.20	416,238	0.20	417,946	0.20	410,038	0.20	413,599	0.20
7	132	−819,695	1,640,899	0.22	**411,609**	**0.18**	**413,101**	**0.19**	408,136	0.22	411,613	0.22
8	151	−807,352	1,616,431	0.20	409,646	0.21	411,260	0.21	402,011	0.21	406,863	0.21
9	170	−790,918	1,583,778	0.23	408,211	0.22	407,848	0.22	405,302	0.25	405,412	0.23

Note. Total sample size equaled 91,648 and subsample size equaled 22,912. For models providing relatively good model-fit results are printed bold faced. BIC = Bayesian information criterion; CE = classification errors.

Table 2. For Each Class of the Five-Class Model, Estimated Class Sizes and l_z^p Mean and Variance Estimates (N = 91,648).

	Class 1 Poor fit		Class 2 Good fit, perfect-SA		Class 3 Mixed-good fit		Class 4 Mixed fit		Class 5 Good fit	
	.10		.25		.20		.25		.21	
	M	Variance	*M*	Variance	*M*	Variance	*M*	Variance	*M*	Variance
CC	−0.54	1.49	0.46	0.44	0.19	0.68	0.27	0.53	0.58	0.26
HA	−0.32	1.01	0.44	0.41	0.24	0.56	0.16	0.57	0.48	0.30
TA	−0.51	1.43	0.35	0.47	0.36	0.37	0.09	0.67	0.47	0.26
FS	−0.27	1.42	0.44	0.28	−0.07	1.12	0.40	0.26	0.47	0.17
RT	−0.46	1.39	0.50	0.42	0.23	0.74	0.19	0.67	0.52	0.31
SA	−0.04	1.05	0.69	0.00	0.01	0.59	0.09	0.14	0.20	0.05
SC	−0.39	1.36	0.41	0.40	0.44	0.34	0.00	0.95	0.41	0.32
ES	−0.10	1.09	0.34	0.35	0.29	0.31	0.24	0.70	0.44	0.21
EC	−0.83	2.62	0.54	0.79	0.62	0.56	−0.24	1.82	0.68	0.49

Note. CC = concentration in class; HA = homework attitude; TA = task attitude; FS = fun at school; RT = relationship with teacher; SA = social acceptance; SC = social competence; ES = expression skills; EC = exam confidence.

Table 3. Coefficients From the Multinomial Logistic Regression and Linear Regression Analysis Predicting Class Membership and l^p_{zm}, Respectively.

	Multinomial logistic regression (MLR)						Linear regression (LR)	
	Class 1 Poor fit	Class 2 Good fit, perfect-SA	Class 3 Mixed-good fit	Class 4 Mixed fit	CE_{Δ} prediction model	R^2_{Δ} pseudo	l^p_{zm} [1]	R^2_{Δ}
Intercept	−1.16	0.03	0.04	−0.25	–	–	−0.20	–
Female (male = ref. group)	−0.20	0.12	−0.06	0.25	.01	.00	0.10	.00
Age	0.00	−0.06	0.24	−0.15	.01	.00	0.07	.00
Education type (effect coding)					.01	.01		.04
Elementary school	1.48	0.22	0.78	0.56	–	–	−1.00	–
Secondary, Level 1	0.76	0.00	0.40	0.33	–	–	−0.55	–
Level 2	−0.34	−0.25	−0.13	−0.17	–	–	0.19	–
Level 3	−0.55	0.03	−0.14	−0.24	–	–	0.43	–
Level 4	−0.64	0.06	−0.42	−0.23	–	–	0.42	–
Level 5	−0.71	0.06	−0.49	−0.27	–	–	0.51	–
Language grade	−0.16	−0.07	−0.04	−0.07	.00	.00	0.14	.00
Completion time	−0.23	0.04	0.06	0.20	.01	.00	0.11	.00
SAQI subscale scores								
Concentration in the Class	−0.18	−0.23	−0.11	−0.25	.01	.00	0.17	.00
Task Attitude	−1.18	−0.03	−0.32	−0.74	.01	.02	0.57	.03
Expression Skills	−1.08	0.16	−0.22	−0.90	.04	.04	1.25	.11
Social Desirability	0.40	0.10	−0.10	0.17	.01	.00	−0.36	.01

Note. Class 5 (good-fit) is used as the reference class. Wald statistics showed a significant overall effects at α = .001 for each predictor in each model. Except for gender and education level, all other explanatory variables were standardized. Level 1 through 5 corresponds to increasing education level: VMBO, VMBO/HAVO, HAVO, HAVO/VWO, and VWO.

[1] l^p_{zm} was standardized.

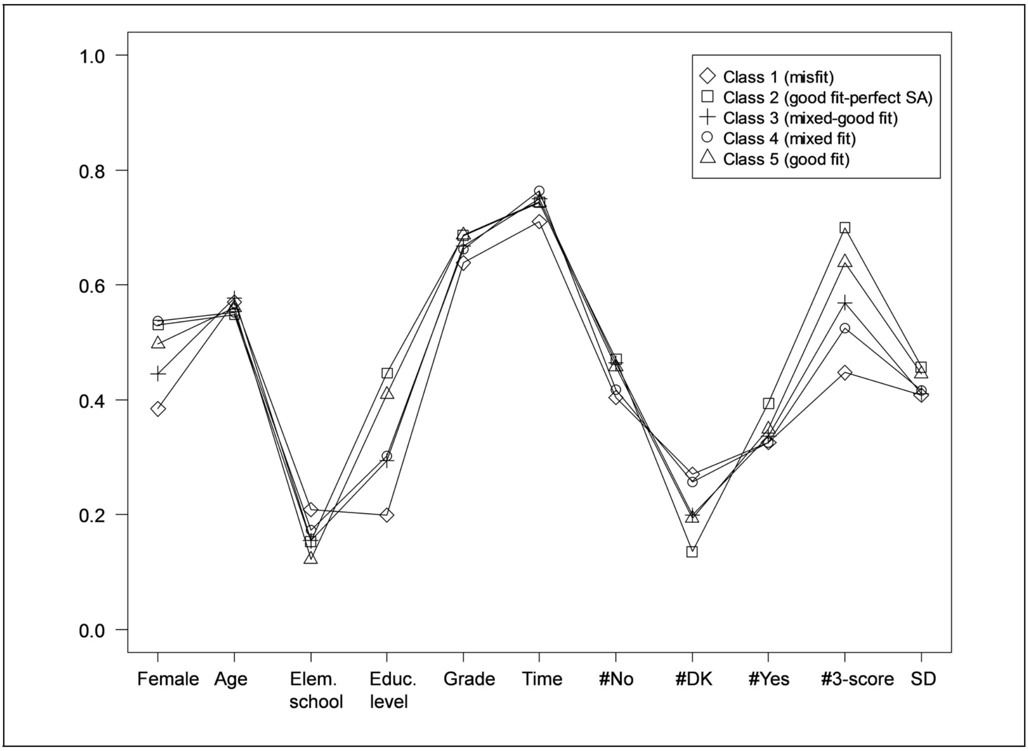

Figure 1. Class-specific mean values for explanatory variables (rescaled to range from 0 to 1).
Note. Rescaling of the class-specific mean values was done by subtracting for each variable the lowest observed variable value from the class-specific variable mean and dividing the result by the range of the observed variable values (i.e., max. value-min. value). The SAQI subscale total scores were not included in the plots because they were strongly related to the 3-score frequency. To represent education type, elementary school (dummy) and education level for secondary school (5 levels) were included in the figure. SAQI = School Attitude Questionnaire Internet. SD= Social desirability bias.

attitude may correspond either to effort or cognitive ability to respond accurately, and hence, they relate negatively to misfit (Krosnick, Narayan, & Smith, 1996); expression skills and language skills (as measured by language grade) are needed to respond accurately, and hence, they are negatively related to misfit (Meijer et al., 2008); and social desirability is a type of aberrant responding, and hence, it is positively related to misfit (Woods, Oltmanns, & Turkheimer, 2008).

Profile plots. Figure 1 shows the profile plot for each of the five LCs. The class-specific means suggest that the classes mainly differ from each other with respect to gender, education level, the frequency of ''don't know'' answers, and the 3-score frequency. Class 1 (poor fit) and Class 2 (good fit, perfect Social Acceptance) have mean covariate values that are most distant from those of the other LCs. Class 1 (poor fit) contains the most boys and elementary school students, and the lowest mean secondary education level, completion time, and 3-score frequency. Class 2 (good fit, perfect Social Acceptance) contains the highest mean education level and a distinct response scale usage: the highest mean number of ''yes'' answers and 3-scores, and the lowest mean number of ''don't know'' answers. Class 5 (good fit) differs from Class 2 (good fit, perfect Social Acceptance) mainly with respect to response scale use. In contrast to the more extreme use of response categories in Class 2, the response scale use of Class 5 (good fit) is close to the average use of each category. Classes 3 (mixed-good fit) and 4 (mixed fit) do not have a distinct profile plot and generally have close to average mean covariate values.

MLR. The explanatory variables improved the separation between classes only little; CE reduced from .173 to .165. CE for the prediction model equaled .60, meaning that 40% of the cases were estimated to be correctly classified. Explained variance as approximated by pseudo R^2 equaled 10%. CE increase due to excluding explanatory variables (Table 2, column 6, CE_Δ) showed that Expression Skills predicted class membership best (4% CE increase) and that excluding most other explanatory variables led to 1% CE increase. Class-specific coefficients larger than .06 in absolute value were significant at $\alpha = .05$. Class 5 (good fit) served as the reference class in MLR. Pseudo R^2 changes showed the same trend as the CE-changes under the prediction model.

Class 5 (good fit) students had highest language grades, concentration, and task attitude across classes. As for education type, students attending elementary school and the lowest secondary school level were the least likely to belong to Class 5 across all classes, whereas children attending Secondary School Level 2 were most likely. Except for the latter effect, results suggest that Class 5 consisted of students that skillfully put effort in responding accurately.

For Class 1 (poor fit), the expected negative age effect was not found, but all other effects on poor fit were as expected. The results were consistent with the authors' expectations; hence, they confirmed that Class 1 is a poor-fit class.

Compared with Class 5, the largest effects for Class 2 (good fit, perfect Social Acceptance) suggested differences in education type and lower concentration scores. Smaller effects suggested that Class 2 students were more often younger girls who responded in a more socially desirable manner, and had higher expression skills and lower language grades. Except for the positive effect of Expression Skills on the probability of Class 2 membership, the other substantial effects suggested Class 2 students to be less skilled to respond accurately than Class 5 students. This explanation is consistent with the results concerning the overall typical-response scale use of Class 2 (Figure 1) and the perfect score on Social Acceptance.

For Class 3 (mixed-good fit), the largest effects indicated that class membership probability increased with age, lower task attitude, and lower expression skills. Also, class membership probability decreased as education level rose. Interpretation of Class 3 was difficult. Compared with Class 1 (poor fit), for Class 4 (mixed fit), the pattern of effects was similar but the effects were weaker. More than Class 1 students, the similar Class 4 students may have put more effort in responding accurately, suggested by the higher completion time (highest across classes).

Comparison of MLR of Class Membership and LR of Continuous l^p_{zm}

The last two columns of Table 3 show results of LR analysis of l^p_{zm}. LR explained 26% of the variance of l^p_{zm}. Effects were as expected, and unique explained variance (R^2_Δ; see Table 3, column 8) was $\geq$ 1% for Education type, Expression Skills, Task Attitude, and Social Desirability. Expression Skills predicted best ($R^2_\Delta = .11$).

MLR and LR detected the same effects. MLR coefficients were inspected more closely. Additional value of MLR can be confirmed when the method identifies explanatory effects deviating from a monotone trend; the presence of such effects implies that LR did not correctly model the relationship between the explanatory variable and person-fit statistic l^p_{zm}. However, when MLR coefficients suggest monotone effects on person-fit class membership, MLR on LC membership did not have additional value other than a possible identification of a nonlinear monotone trend.

MLR coefficients confirmed a monotone relationship between Language grade, Task Attitude, and, apart from some small deviations, Education type and person fit. For example, the LR effect of Task Attitude was positive, and consistent with this effect, the MLR model coefficients suggested that the well-fitting classes (Classes 2 and 5) had the highest Task Attitude, followed by Class 3 (mixed-good fit), Class 4 (mixed fit), and Class 1 (poor fit). Hence, MLR of class membership added little information about the relationship between these explanatory variables and person fit. For these explanatory variables, LR may be more appropriate than MLR because the continuous person-fit statistic can reveal the explanatory effects well and categorization of respondents into latent person-fit classes is unnecessary.

For most other explanatory variables, MLR showed patterns of effects that are inconsistent with a monotone trend. For example, LR suggested that higher completion time produces better person fit. Consistent with this result, MLR showed that completion time is the shortest in the poor-fit Class 1 but it also showed that completion time is longer in the mixed-fit Class 4 than in the good-fit Classes 2 and 5. This result suggests that some students (i.e., Class 4) showed some misfit although they spent more time and effort answering questions than the well-fitting students while others showing substantial misfit (Class 1) spent little time answering questions. Hence, in contrast to LR, MLR suggested absence of person fit is not only due to rushing through the questions but requires a more involved explanation.

Concentration in the Class provides another example of the additional value of MLR of class membership. LR suggested that higher concentration led to better person fit. Consistent with this result, MLR showed that as expected the good-fit Class 5 (baseline) had the highest concentration. However, the MLR effects also showed that, contrary to expectation, not the poorly fitting students in Class 1 but the well-fitting students in Class 2 and those in the mixed-fit class had the lowest concentration scores. This implies that good concentration is not required for every type of good fit but only for the type represented by Class 5. Possibly, combined with the socially desirable response style already identified for Class 2, it may be concluded that little concentration and effort does not necessarily interfere with good fit. Further inspection of MLR effects showed that gender and age effects in MLR also differ from a monotone pattern but space limitations prevent further discussion. However, it is noted that the relative value of the explanatory variables in the model reflected by CE_{Δ} and R^2_{Δ} also suggests that MLR captures these effects better than LR: Compared with the other explanatory variables, Concentration in the Class, Completion Time, Gender, and Age are relatively more important in MLR than in LR.

Finally, MLR effects of Education type, Expression Skills, and Social Desirability mostly showed a monotone pattern. However, the effects that these explanatory variables have on the probability to belong to Class 2 compared with Class 5 suggest various interesting differences between the two classes for which l^p_{zm} classification results in a person misfit prevalence of zero. To conclude, MLR of class membership showed more complete insight into the different causes of good fit and poor fit than LR of continuous person-fit statistic l^p_{zm}.

Discussion

The authors evaluated whether a novel approach to PFA based on LC cluster analysis of subscale l^p_z person-fit statistics contributes to a better understanding of aberrant responding compared with treating person fit on multi-scale measures as a single continuous variable. The LC cluster analysis of person fit to the nine SAQI subscales resulted in a five-class model. As expected, the five-class model included a poor-fit class and a good-fit class. Contrary to what was expected, the classes of mixed person fit (Classes 3 and 4) did not show misfit on subscales with related content; only one subscale in each class provided evidence of (mild) misfit. Cross-validation suggested that the five-class model was unstable. However, similar l_z^p class profiles were found across the models identified in different validation subsamples, suggesting that LC cluster analysis provided robust information for explanatory PFA. LC cluster analyses showed that person misfit quantified by the l_z^p statistic may provide a simplistic representation of good fit and poor fit and thus requires a more in-depth study.

With the exception of Class 1 (poor fit), the authors were unable to interpret and distinguish the other person-fit classes well only on the basis of their estimated l_z^p mean and variance structure. MLR of class membership provided meaning to most latent person-fit classes. In particular, the demographic and response style variables were informative about class distinction. Students in good-fit Class 2 appeared to be more inclined to socially desirable responding, whereas students in good-fit Class 5 made normal use of response categories. Classes 3 and 4 both showed a mixed pattern of l_z^p means, uninformative of consistent good fit or poor fit. MLR showed that Class 4 (mixed fit) was similar to Class 1 (poor fit) with respect to background characteristics (e.g., low education level, low school task attitude) and therefore may have been prone to some degree of misfit. Class 3 was most difficult to interpret; the l_z^p structure of this class appeared normal, supporting this conclusion.

MLR of class membership revealed a non-monotone relation between several explanatory variables and the categorical dependent person-fit variable. Thus, it was concluded that LR of continuous person-fit statistic l_{zm}^p, thus assuming linearity and monotonicity, fell short in this data-example. Different processes may cause person fit and person misfit, resulting in multiple good fit and multiple misfit classes classes providing information not attainable by means of LR. For example, the present study suggested that good person fit may be due to different response processes, one of which may be socially desirable responding. Some students may have had less trouble consistently selecting the most socially desirable response category than responding on the basis of their true trait value, resulting in good fit due to social desirability bias. Likewise, Ferrando (2014) suggested that ''overconsistent'' responding to an extraversion scale may reflect response styles. Stukenberg, Brady, and Klinetob (2000) suggested that overconsistent responding to the Minnesota Multiphasic Personality Inventory (MMPI) could identify exaggerated protocols in clinical populations. Inconsistent with these results, the present study found the Social Desirability scale of the SAQI to be only weakly related to class membership. More research focusing on causes and implications of overconsistency is needed.

LC-PFA may be a viable alternative for explanatory PFA, but based on results of this study showing instability of the LC-PFA cluster solution, the usage of LC-PFA approach for making decisions about individual person-fit and misfit classification in real-data applications is discouraged. Instead, alternative multiscale person-fit approaches are recommended to detect person misfit on questionnaires consisting of multiple subscales (e.g., Conijn et al., 2014; Drasgow et al., 1991). Explanations for the instability of the LC solution are unreliability of the subscale person-fit statistics based on few items and heterogeneity of the sample; in the present study, GRM parameters may have differed between school levels and resulting lack of measurement invariance may have caused person misfit. Simulation studies may address the circumstances in which LC-PFA classifies students correctly. Also, such simulations could be used to determine whether explanatory variables for class membership are correctly identified despite possible unreliable classification of individual respondents.

To conclude, future research may apply LC cluster analysis of person-fit data to multiscale measures containing longer subscales and more homogeneous samples and address individual-level classification. The aim of the current study was to investigate the potential of LC-PFA for explaining aberrant responding to multiscale measures. Based on a diverse set of analyses, the authors consistently found that LC cluster analysis of person-fit statistics based on related short scales revealed important information about responding that would remain unknown had only a simple dichotomous or continuous person-fit statistic been used.

Acknowledgment

The authors thank Harrie C. M. Vorst for providing the data of the School Attitude Questionnaire Internet.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Conijn, J. M., Emons, W. H. M., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire-45. *Assessment*, *22*, 513-524.

Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic l_z based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*, 122-136.

Conijn, J. M., Emons, W. H. M., Van Assen, M. A. L. M., Pedersen, S. S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research*, *48*, 692-718.

De la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159-177.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person response functions. *Psychological Methods*, *10*, 101-119.

Ferrando, P. J. (2014). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement*, *38*, 166-183.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.

Krosnick, J. A., Narayan, S. S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research* (pp. 29-44). San Francisco, CA: Jossey-Bass.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception profile for children. *Journal of Personality Assessment*, *90*, 227-238.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Muthén, B. O., & Muthén, L. K. (2007). *Mplus: Statistical analysis with latent variables* (Version 5.0). Los Angeles, CA: Statmodel.

Pinsoneault, T. B. (2002). A variable response inconsistency scale and a true response inconsistency scale for the Millon Adolescent Clinical Inventory. *Psychological Assessment*, *14*, 320-328.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Stukenberg, K., Brady, C., & Klinetob, N. (2000). Use of the MMPI-2's VRIN scale with severely disturbed populations: Consistent responding may be more problematic than inconsistent responding. *Psychological Reports*, *86*, 3-14.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Vermunt, J. K., & Magidson, J. (2005). *Latent gold 4.0 user's guide*. Belmont, MA: Statistical Innovations.

Vorst, H. C. M., Smits, J. A. E., Oort, F. J., Stouthard, M. E. A., & David, S. A. (2008). *Schoolvragenlijst voor basisonderwijs en Voortgezet Onderwijs* [School Attitude Questionnaire Internet for elementary education and high school education]. Amsterdam, The Netherlands: Pearson Assessment and Information BV.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159-168.