Sensitivity Analyses for the Mixed Coefficients Multinomial Logit Model

Daniel Kasper, Ali Ünlü, and Bernhard Gschrey

Abstract For scaling items and persons in large scale assessment studies such as Programme for International Student Assessment (PISA; OECD, PISA 2009 Technical Report. OECD Publishing, Paris, 2012) or Progress in International Reading Literacy Study (PIRLS; Martin et al., PIRLS 2006 Technical Report. TIMSS & PIRLS International Study Center, Chestnut Hill, 2007) variants of the Rasch model (Fischer and Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications. Springer, New York, 1995) are used. However, goodness-of-fit statistics for the overall fit of the models under varying conditions as well as specific statistics for the various testable consequences of the models (Steyer and Eid, Messen und Testen [Measuring and Testing]. Springer, Berlin, 2001) are rarely, if at all, presented in the published reports.

In this paper, we apply the mixed coefficients multinomial logit model (Adams et al., The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1–23, 1997) to PISA data under varying conditions for dealing with missing data. On the basis of various overall and specific fit statistics, we compare how sensitive this model is, across changing conditions. The results of our study will help in quantifying how meaningful the findings from large scale assessment studies can be. In particular, we report that the proportion of missing values and the mechanism behind missingness are relevant factors for estimation accuracy, and that imputing missing values in large scale assessment settings may not lead to more precise results.

D. Kasper · A. Ünlü (✉) · B. Gschrey
Chair for Methods in Empirical Educational Research, TUM School of Education, and Centre for International Student Assessment (ZIB), Technische Universität München, Arcisstrasse 21, 80333 Munich, Germany
e-mail: daniel.kasper@tum.de; ali.uenlue@tum.de; bernhard.gschrey@tum.de

M. Spiliopoulou et al. (eds.), *Data Analysis, Machine Learning and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-01595-8_42,

1 Introduction

To analyze data obtained from large scale assessment studies such as Programme for International Student Assessment (PISA; OECD 2012) or Progress in International Reading Literacy Study (PIRLS; Martin et al. 2007) different versions of the Rasch model (Fischer and Molenaar 1995) are applied. For instance, in PISA the mixed coefficients multinomial logit model (Adams et al. 1997) has been established, to scale items and persons. One may be interested in how well the model fits the data. But goodness-of-fit statistics for the overall fit of this model under varying conditions are rarely presented in the published reports, if they are presented at all (OECD 2002, 2005, 2009, 2012).

One special characteristic of the PISA assessment data is the presence of missing values. Missing values can occur due to missing by design as well as item non-response (OECD 2012). The handling of missing values seems to be crucial, because an improper treatment of missing values may result in invalid statistical inferences (Huisman and Molenaar 2001). In this paper, we apply the mixed coefficients multinomial logit model to PISA data for varying forms of appearance of missing values. Based on various overall and specific fit statistics, we compare how sensitive this model is, across changing conditions.

In the mixed coefficients multinomial logit model, the items are described by a fixed set of unknown parameters, $\boldsymbol{\xi}$, and the student outcome levels (the latent variable), $\boldsymbol{\theta}$, are random effects. Three parts of this models can be distinguished: the conditional item response model, the population model, and the unconditional, or marginal, item response model (for technical details, see Adams et al. 1997).

In addition to the afore mentioned components, a posterior distribution for the latent variable for each individual n is specified by

$$h_\theta(\boldsymbol{\theta}_n; \boldsymbol{w}_n, \boldsymbol{\xi}, \boldsymbol{\Gamma}, \Sigma | \boldsymbol{x}_n) = \frac{f_x(\boldsymbol{x}_n; \boldsymbol{\xi} | \boldsymbol{\theta}_n) f_\theta(\boldsymbol{\theta}_n)}{\int_{\boldsymbol{\theta}} f_x(\boldsymbol{x}_n; \boldsymbol{\xi} | \boldsymbol{\theta}) f_\theta(\boldsymbol{\theta})},$$

where $\boldsymbol{x}_n$ is the response vector, and $\boldsymbol{\Gamma}$, $\boldsymbol{w}_n$, and $\boldsymbol{\Sigma}$ are parametrizing the postulated multivariate normal distribution for $\boldsymbol{\theta}$ (OECD 2012, p. 131). Estimates for $\boldsymbol{\theta}$ are random draws from this posterior distribution for each student, and these are referred to as plausible values (see Mislevy 1991; Mislevy et al. 1992).

The mixed coefficients multinomial logit model is used in PISA for three purposes: national calibration, international scaling, and student score generation (estimation of students' plausible values). Multidimensional versions of this model have been fitted to PISA data; for instance, a three-dimensional version has had reading, science, and mathematics as its (correlated) dimensions. For estimating the parameters of this model, the software *ConQuest*® can be used (Wu et al. 2007).

Missing values in PISA can occur due to missing by design (different students are administered different test items) as well as by item non-response. Usually three mechanisms producing item non-response are distinguished: Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random

(NMAR) (Little and Rubin 2002; Schafer 1997). When MCAR, missing item scores form a simple random sample from all scores in the data, that is, there is no relation to the value of the item score that is missing, or to any other variable. If missingness is related to one or more observed variables in the data, the process is called MAR. NMAR means that missingness is related to the value that is missing or to unobserved variables.

To control for item non-response, different procedures are studied in the statistical literature (Huisman and Molenaar 2001). One popular technique is *imputation*. Using this technique, missing responses are estimated, and the estimates are substituted for the missing entries. However, a number of imputation techniques are available (e.g., see van Ginkel et al. 2007a), so the question is what methods are to be preferred.

Huisman and Molenaar (2001) compared six imputation methods for dealing with missing values. They used four real complete data sets with different sample sizes, and missing values were created in the samples using three different mechanisms resulting in MCAR, MAR, and NMAR. The proportion of created missing values was $P = 0.05$, $P = 0.10$, and $P = 0.20$. In general, model based imputation techniques perform better than randomization approaches. But this effect can only be observed for a missing value proportion of at least $P = 0.10$ and when missingness is due to MAR or NMAR. An effect due to sample size could not be observed.

Van Ginkel et al. (2010) used two-way imputation with error and compared it with listwise deletion. The method of two-way imputation is based on a two-way ANOVA model. It produces relatively unbiased results regarding such measures as Cronbach's alpha, the mean of squares in ANOVA, item means, mean test score, or the loadings from principal components analysis. A description of the two-way ANOVA model can be found in van Ginkel et al. (2007c). Missingness was introduced into a real complete data set using the mechanisms MCAR, MAR, and NMAR. The data set consisted of ten unidimensional items. The method of two-way imputation with error outperformed listwise deletion with respect to different criteria (e.g., Cronbach's alpha and mean test score). The results were almost as good as those obtained from the complete data set.

The strength of the method of two-way imputation with error (TW+e) was also shown in several other studies (van der Ark and Sijtsma 2005; van Ginkel et al. 2007b, 2007c). This method may also be useful for large scale assessment studies such as PISA. Another imputation method considered is multiple imputation by chained equations (MICE), which is a multiple imputation technique that operates in the context of regression models. In general, missing values are replaced by plausible substitutes based on the distribution of the data. The MICE procedure contains a series of regression models, where each variable with missing data is modeled conditional on the other variables in the data. Iterations then yield multiple imputations (for a detailed explanation of the method, see Azur et al. 2011).

Because of large number of variables (more than 200) and respondents (around half a million) sophisticated methods of imputation such as the multiple imputation by chained equations (van Buuren et al. 2006; van Buuren 2007) possibly may not

be applicable. Unfortunately, information about how different methods for dealing with missing values perform in the context of PISA are lacking so far. In this regard, the present paper will study whether the application of these imputation methods may lead to improved estimates. The afore mentioned studies can only serve as a reference, for how sensitive the mixed coefficients multinomial logit model may "react" to missing values or different imputation methods. The reason for this is, that none of the studies have investigated the sensitivity of multidimensional versions of Rasch type models for missing value analyses. Moreover, crucial criteria such as the goodness-of-fit of these models or the accuracy of the item parameter estimates have not been investigated in those studies.

2 Study Design

To study the estimation accuracy of the mixed coefficients multinomial logit model under varying conditions for missing values, we analyzed data from the PISA 2009 study (OECD 2012). We used a complete data set of 338 German students on the mathematics and science test items of Booklet Nr. 9. Missing values for this data set were created using the mechanisms MCAR, MAR, and NMAR. For MCAR, each data point had the same probability of being coded as missing value. Under the condition of MAR, missingness was associated with gender: for men, the probability of a missing value was nine times higher than for women. To reach NMAR, in addition to the correlation of missingness with gender, the probability of a missing value was eight times higher for incorrect answers (that is, for zero entries) than for correct answers.

Three proportions of missing values were considered: $P = 0.01$, $P = 0.03$, and $P = 0.05$. These proportions capture the usual amount of missingness in the PISA test booklets. As imputation methods, we used two-way imputation with error (TW+e) and multiple imputation by chained equations (MICE). Each of the imputation methods was applied one time to every data set, so for any imputation method, missing condition, and proportion of missing values, there is one imputed data set.

All of the $2 \times 3 \times 3 \times 1$ imputed data sets, the nine missing data sets (MD), and the complete data set were analyzed with the mixed coefficients multinomial logit model, whereat the mathematical items were allocated to one dimension and the science items to another dimension. As criteria for the sensitivity of this model, the item fit statistic MNSQ and the item parameter estimates were used. MNSQ quantifies how well the model fits the data. This fit statistic is applicable especially for large numbers of observations. A perfect value of MNSQ is 1.0, whereas values less than 1.0 indicate an overfit, values greater than 1.0 an underfit. In general, mean squares in a near vicinity of 1.0 indicate little distortion. On the other hand, the item parameters may be interpreted as the difficulties or discrimination intensities of the items, and theoretically, they can range in the reals or subsets

thereof. As parameters of the mixed coefficients multinomial logit model, they can be estimated by maximum likelihood procedures.

For both statistics, we calculated the differences between the estimates obtained from the complete data sets and the estimates for the missing values and imputed data sets. The absolute values of these differences were averaged and the standard deviations were calculated. In addition, ANOVA models were applied.

3 Results

The means of the absolute differences for MNSQ between the estimates from the complete data sets and the estimates for the missingness and imputed data sets are summarized in Table 1. As can be seen, the mean differences in MNSQ between the complete data sets and the imputed as well as the missing data sets are small. As expected, the difference is larger when the proportion of missing values increases. The mechanisms behind missingness obviously influence the estimation accuracy in the case of NMAR. The effect of imputation methods on estimation accuracy is small. In general, using the missing data set (MD) for the analysis results in the least biased estimates.

As the results of the ANOVA show, the small effects of imputation methods on estimation accuracy in terms of MNSQ are statistically significant (Table 2). Also the effects of the proportion of missing values and NMAR on the estimation accuracy in terms of MNSQ are statistically significant. In addition, all two-way interaction terms were included in the model, but were not significant.

The means of the absolute differences for the estimated item parameters between the complete data sets and the missingness and imputed data sets are summarized in Table 3. Generally, these results are similar to the previous findings. We observe small differences between the estimates obtained from the complete data sets and the imputed as well as the missing data sets. The difference is larger when the proportion of missing values increases, and an effect of the mechanisms underlying missingness can be observed for NMAR.

As the results of the ANOVA show, the effects of the proportion of missing values ($P = 0.05$), NMAR, TW+e, and MICE on the estimation accuracy in terms of the item parameter estimates are statistically significant (Table 4). In addition, all two-way interaction terms were included in the model, but were not significant.

4 Discussion

For scaling items and persons in PISA, the mixed coefficients multinomial logit model is used. However, statistics for the fit of this model under varying conditions for dealing with missing values are rarely, if at all, presented in the published reports. We have applied the mixed coefficients multinomial logit model to PISA

Table 1 Means of the absolute differences for MNSQ

	MCAR			MAR			NMAR		
	P			*P*			*P*		
Imputation Method	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05
TW+e	0.008	0.005	0.012	0.006	0.011	0.013	0.008	0.013	0.015
MICE	0.007	0.008	0.011	0.006	0.013	0.014	0.006	0.014	0.018
MD	0.006	0.007	0.009	0.006	0.009	0.012	0.004	0.011	0.011

TW+e two-way imputation with error, *MICE* multiple imputation by chained equations, *MD* missing data set (no imputation)

Table 2 ANOVA for mean difference for MNSQ

Effect	*F*	*df*1	*df*2	*p*
TW+e[a]	4.51	1	795	0.03
MICE[a]	5.60	1	795	0.02
P3[b]	0.83	1	795	0.36
P5[b]	75.51	1	795	0.00
MAR[c]	1.00	1	795	0.32
NMAR[c]	16.80	1	795	0.00
TW+e*MAR	0.59	1	795	0.44
MICE*MAR	0.00	1	795	0.97
TW+e*NMAR	1.18	1	795	0.28
MICE*NMAR	1.41	1	795	0.24
TW+e*P3	1.17	1	795	0.28
TW+e*P5	0.09	1	795	0.76
MICE*P3	0.56	1	795	0.46
MICE*P5	2.93	1	795	0.09

[a] Reference category was no imputation
[b] Reference category was $P = 0.01$
[c] Reference category was MCAR

Table 3 Means of the absolute differences for estimated item parameters

	MCAR			MAR			NMAR		
	P			*P*			*P*		
Imputation Method	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05
TW+e	0.013	0.018	0.035	0.012	0.028	0.035	0.017	0.032	0.045
MICE	0.011	0.024	0.031	0.011	0.020	0.045	0.015	0.034	0.043
MD	0.008	0.014	0.018	0.009	0.023	0.029	0.012	0.019	0.025

TW+e two-way imputation with error, *MICE* multiple imputation by chained equations, *MD* missing data set (no imputation)

data under varying conditions for missing values. Based on various fit statistics, we have compared how sensitive this model is, across changing conditions.

With respect to the fit criterion MNSQ, we have shown that the proportion of missing values obviously influences estimation accuracy; less accurate estimates are observed for higher proportions of missing values. The mechanisms behind

Table 4 ANOVA for mean difference for estimated item parameters

Effect	F	$df1$	$df2$	p
TW+e[a]	19.39	1	741	0.00
MICE[a]	5.87	1	741	0.02
P3[b]	0.14	1	741	0.72
P5[b]	126.36	1	741	0.00
MAR[c]	0.07	1	741	0.78
NMAR[c]	16.59	1	741	0.00
TW+e*MAR	2.12	1	741	0.15
MICE*MAR	0.55	1	741	0.46
TW+e*NMAR	0.83	1	741	0.36
MICE*NMAR	0.06	1	741	0.80
TW+e*P3	0.27	1	741	0.60
TW+e*P5	3.69	1	741	0.06
MICE*P3	0.05	1	741	0.83
MICE*P5	3.12	1	741	0.08

[a] Reference category was no imputation
[b] Reference category was $P = 0.01$
[c] Reference category was MCAR

missingness also appear to be relevant for estimation accuracy. As the study of this paper corroborates, imputing missing values does not lead to more precise results in general. In future research, it would be interesting to investigate the effects of imputation techniques in matters of higher proportions of missing values, as well as of appropriate modifications of the mixed coefficients multinomial logit model for the lower proportions in PISA.

Generally, the pattern of results for the estimated item parameters resembles the results for MNSQ. Again, the proportion of missingness, the imputation methods, and the mechanisms creating the missing values have an influence on the estimation accuracy. It seems that the imputation methods considered here do not lead to more accurate results regarding the fit criterion and the item parameters, at least under the conditions studied in this paper.

Which of the imputation techniques should be preferred in educational large scale assessment studies such as PISA? The findings of this paper cannot favor one over the other, of the two analyzed imputation techniques MICE and TW+e. Similar results were obtained for both methods. In some cases, the TW+e method led to better results, and in other cases, MICE performed better.

Nonetheless, the considered missingness proportions were relatively small, and the investigation of the influence of missing values on such other criteria as the important students' plausible values in PISA would have exceeded the scope of this paper. These topics must be pursued in future research.

References

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*, 40–49.

Fischer, G. H., & Molenaar, I. W. (Eds.), (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York: Springer.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data*. New York: Wiley.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report*. Chestnut Hill: TIMSS & PIRLS International Study Center.

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

OECD (2002). *PISA 2000 Technical Report*. Paris: OECD Publishing.

OECD (2005). *PISA 2003 Technical Report*. Paris: OECD Publishing.

OECD (2009). *PISA 2006 Technical Report*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Steyer, R., & Eid, M. (2001). *Messen und Testen* [*Measuring and Testing*]. Berlin: Springer.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*, 219–242.

Van Buuren, S., Brand, J., Groothuis-Oudshoorn, C., & Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049–1064.

Van der Ark, L. A., & Sijtsma, K. (2005). The effect of missing data imputation on Mokken scale analysis. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 147–166). Mahwah: Erlbaum.

Van Ginkel, J., Sijtsma, K., Van der Ark, L. A., & Vermunt, J. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology, 6*, 17–30.

Van Ginkel, J., Van der Ark, L. A., & Sijtsma, K. (2007a). Multiple imputation for item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*, 387–414.

Van Ginkel, J., Van der Ark, L. A., & Sijtsma, K. (2007b). Multiple Imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60*, 315–337.

Van Ginkel, J., Van der Ark, L. A., Sijtsma, K., & Vermunt, J. (2007c). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis, 51*, 4013–4027.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER ConQuest: Generalised item response modelling software*. Camberwell: ACER Press.