Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Chapter 9

Discussion

9.1 Introduction

This thesis focused on generalizing nonparametric item response theory (IRT) and Mokken scale analysis (MSA) to multilevel test data (i.e., data consisting of item scores of respondents nested in clusters). Nonparametric IRT models are especially useful for ordinal measurement, as part of theory development of psychological constructs, or for psychological constructs for which creating a large number of items is difficult (Mokken, 1971, pp. 115–116; Sijtsma & Van der Ark, 2020, pp. 107–109). MSA provides a selection of methods for constructing scales and evaluating model fit based on on nonparametric IRT (e.g., Mokken, 1971; Sijtsma & Van der Ark, 2017). There were two main reasons why MSA was not suitable for multilevel test data. First, methods in MSA assume that item scores are obtained from respondents using simple random sampling, which is violated in multilevel test data. Ignoring the multilevel structure may lead to the inclusion of items to the test that do not contribute to (or possibly negatively affect) accurate measurement, or the quality of the test may be overestimated. Second, methods in MSA provide results on the respondent level only, not on the group level. Hence, for multilevel test data, currently available MSA methods are only of limited value.

This thesis contributed to making MSA suitable to multilevel test data and improved some existing methods in MSA for one-level data in the following ways. First, by discussing solutions to two computational problems pertaining Guttman errors, relevant for estimating Mokken's and two-level scalability coefficients (Chapter 2). Second, by deriving and optimizing point and interval estimates of two-level scalability coefficients in multi-rater data (Chapters 3 and 4) and of Mokken's (1971) scalability coefficients in clustered data (Chapters 5 and 6). Third, by introducing a test-guided automated item selection procedure (T-AISP) that is available for both nonclustered and clustered test scores (Chapter 7). Finally, by proposing four two-level nonparametric IRT models (two respondent-level models and two group-level models), with their implied measurement and data properties (Chapter 8). These developments contribute to knowledge on how tests and questionnaires can be investigated in the presence of multilevel test data, hopefully

leading to improved measurement of psychological constructs of nested respondents, or of subjects (groups or persons) that are scored by multiple raters.

9.2 Main Findings

Summarizing results across chapters led to the following general conclusions. Point estimates of the two-level scalability coefficients were unbiased in all conditions. In general, the *two-level method* for estimating standard errors (Chapters 3 and 6) resulted in unbiased standard error estimates for all types of scalability coefficients in two-level data, and can therefore be confidently used in practice, preferably for scales of at least four items. Note that, initially, these standard errors were biased for samples with an unequal number of respondents per group (Chapter 4), but this problem was solved in Chapter 6. However, sufficient independent information needs to be present in the data set to get accurate estimates (see also, Snijders & Bosker, 2012, p. 24). For example, for samples consisting of only a few large groups with substantial within-group dependency, the two-level standard error estimates were negatively biased and resulted in undercoverage of the confidence intervals (Chapter 4). Standard errors were derived by assuming that the item-score pattern frequencies follow a multinomial distribution for each group. Alternative distributions (e.g., a Dirichlet-multinomial distribution) are computationally more complex, but may give better results in situations where standard errors were less accurate.

Because the distribution of the two-level scalability coefficients is asymptotically normal, initially Wald-based confidence intervals were used for confidence interval construction (Chapter 3). However, the Wald-based interval may be biased near the boundary of 1, hence, an alternative, range-preserving confidence interval was introduced in Chapter 5. Results from a simulation study showed that, in one-level (i.e., nonclustered) data, for Mokken's scalability coefficients not close to 1, Wald-based and range-preserving methods were very similar and both useful. For very strong scales ($H > .7$), the range-preserving methods were more accurate and preferred over the Wald-based method, especially for the left tail of the sampling distribution (used in one-sided significance tests). In Chapter 6, however, for clustered data, the confidence interval of Mokken's scalability coefficient was somewhat asymmetric, possibly caused by skewness of the sampling distribution. As the symmetry was slightly better for the Wald-based confidence interval, we suggested using a Wald-based interval if the scalability coefficient is (likely) well below the upper boundary of 1. As in practice, scalability coefficients are seldom very close to 1, and usually hypotheses regarding scalability coefficients are in the range 0 to .55, the Wald-based confidence intervals and significance tests were recommended for general use.

Chapter 7 pointed out that the automated item selection procedure (AISP) in MSA insufficiently takes sampling fluctuation of Mokken's scalability coefficients into account. Hence, a Wald-based significance test was implemented in the AISP algorithm, which resulted in a test-guided AISP (T-AISP), that is available for both nonclustered and

clustered test scores. By definition, the T-AISP is more conservative than the AISP. So when using the T-AISP, a researcher may expect that less items will end up in the scale. However, the final scale significantly complies with the criteria of a Mokken scale. The T-AISP was integrated into a two-step, test-guided MSA for scale construction, to guide the analysis for nonclustered and clustered data.

Chapter 8 introduced four two-level nonparametric IRT models, two respondent-level models and two group-level models. These models are generalizations of two existing nonparametric IRT models: the monotone homogeneity model (MHM) and the double monotonicity model (DMM). The MHM-1 and the DMM-1 assume unidimensionality, local independence, monotonicity, and, for DMM only, an invariant item ordering on the respondent-level item scores. The MHM-2 and the DMM-2 assume unidimensionality, local independence, monotonicity, and, for DMM only, an invariant item ordering on the group-level item scores. Their hierarchical relation shows that DMM-1 implies all other models and that MHM-2 is the most general model. Originally, the DMM defined was in terms of nonintersecting item-step response functions (i.e., all item-score categories are ordered), rather than in terms of nonintersecting item-response functions (Molenaar, 1997). As investigating properties of items can be considered more relevant than investigating properties of item-steps, the new definition of the DMM can be considered more useful. If there is reason to require an invariant item-step order, an alternative DMM-like model may be proposed including this assumption. However, one should realize that an invariant item-step order not necessarily implies an invariant item order (Sijtsma & Hemker, 1998).

For each two-level model, ordering properties were derived that justify stochastic ordering of respondents on the latent variable using the observed respondent-level test score, the ordering of groups on the latent variable using the observed group-level test score, and/or the ordering of items using the observed mean item score. In addition, we derived observable data properties implied by the models, which can be used to investigate the model fit for a given data set. Specifically, we generalized the properties manifest monotonicity, conditional association, and manifest invariant item ordering for the respondent level and the group level.

Two-level nonparametric IRT models were defined on either or both the respondent level and the group level. Depending on the interest of the researcher, one or both levels are relevant for scaling. If the goal is scaling respondents, Mokken's scalability coefficients and the respondent-level assumptions of the MHM-1 or DMM-1 are of key interest. If the goal is scaling groups, as is the case in multi-rater data, the two-level scalability coefficients and group-level assumptions of MHM-1, DMM-1, MHM-2 or DMM-2 are of key interest. For example, if a group-level IRF is flat, an item does not discriminate between low and high values of γ. Such an item does not contribute to accurate measurement of the group. However, the respondent-level assumptions may still prove informative for investigating, for example, whether the respondents may also be ordered using their sum score or how the results compare across levels. Therefore, investigating the MHM-1 or DMM-1 by using methods on both the respondent level and the group level was suggested. If the analysis

on both levels is compared, it should be realized that item and test scores may have a different interpretation in terms of the psychological construct (Stapleton et al., 2016). If an analysis on both levels is not possible (e.g., only group-level data are available), one should investigate whether the MHM-2 or the DMM-2 is an appropriate model.

The computational developments throughout this thesis were implemented in R (R Core Team, 2020) in the package `mokken` (Van der Ark, 2007, 2012). The function `MLweight()` gives the estimated Guttman errors in two-level data, based on the insights from Chapter 2. Function `MLcoefH()` gives estimates for the two-level scalability coefficient, their standard errors, and, optionally, Wald-based or range-preserving confidence intervals. The existing function `coefH()`, used for estimating Mokken's scalability coefficients and their standard errors, was updated to also allow clustered data and, optionally, give Wald-based or range-preserving confidence intervals. Function `coefZ()` was updated to give the Wald-based or the range-preserving test statistics for Mokken's scalability coefficients, available for both nonclustered or clustered data, and function `MLcoefZ()` was introduced to give these test statistics for two-level scalability coefficients. Function `aisp()` was adjusted to also allow clustered data, and to include the T-AISP algorithm. Finally, function `ICC()` computes and tests the intraclass correlation for two-level data.

9.3 Future Research Directions

The research in this thesis expanded on the model by Snijders (2001a), because of its strong link to Mokken's MHM and DMM and, therefore, to MSA. However, several generalizations of the MHM and DMM are possible. Within the framework proposed in Chapter 8, one may also consider a model that orders respondents only within a group. For this model, the IRFs are assumed to be increasing only in δ. Properties and applications of this model are yet unknown. Outside the framework proposed in Chapter 8, in Chapter 4, the nonparametric hierarchical rater model was proposed, a nonparametric version of the (parametric) hierarchical rater model (Patz et al., 2002). Possibly other two-level parametric IRT models may be redefined as a nonparametric model, such as the multiple raters model (Verhelst & Verstralen, 2001) or the rater bundle model (Wilson & Hoskens, 2001). Alternatively, the existing nonparametric partial credit model or nonparametric sequential model (Hemker et al., 1997, 2001, respectively) may be generalized to a two-level framework. In addition, multidimensional generalizations may be useful for developing scales that explicitly have a separate respondent and group component. How these alternative models hierarchically relate to the models presented in Chapter 8, and what properties they imply, is a topic for further investigation.

The proposed methods enabled investigating a group-level scale using a fixed set of items. However, group-level test construction would benefit from the development of a group-level T-AISP. There are three classes of two-level scalability coefficients (i.e., within-rater, between-rater, and ratio coefficients), and future research may focus on how these coefficients can be optimally implemented in an algorithm for selecting a subset of

items from a larger set of items. Furthermore, the model fit methods derived in Chapter 8 pertain to the population level. Existing methods for one-level data divide respondents in item- or rest-score groups to test violations of the observable properties (e.g., Ligtvoet et al., 2010; Molenaar & Sijtsma, 2000; Straat et al., 2016). Currently, it is unclear if or how within-group dependency in two-level data affect statistical analysis after dividing respondents in rest-score groups. Future research should investigate whether the existing computational methods and statistical tests may still give accurate results or develop methods for two-level test data.

The proposed methods assume that respondents are stochastically independent given their group-membership, or, statistically, that the individual respondent components are independent and identically distributed. This assumption may be considered realistic if the data are obtained by a clustered or multi-stage sampling design. However, it is possible to imagine scenarios in which this assumption may be less realistic; for example, if respondents have fixed roles (e.g., for each child in the sample, its parent, its teacher, and its therapist each score the items on a test). Whether or how the developed models and methods are useful in such situations is a topic for further investigations. In line with this suggestion, the procedures would benefit from generalizations to more complex sampling designs, such as a three-level nested sample, for example if students are nested in classrooms, nested in schools, or a cross-nested sample, where raters score multiple groups. In the latter case it may also be possible to investigate bias and consistency of individual raters.

Note that the application of two-level scalability coefficients and their standard errors is not limited to multi-rater data. They may also be applied in research with multiple (random) circumstances or time points in which the same questionnaire is completed. Also, the items may be replaced by a fixed set of situations in which a particular skill is scored using a single item. In addition, if the scalability of a multi-rater test is deemed satisfactory, a related (but different) topic concerns the reliability. For a given test, Snijders (2001a, p. 13) presented coefficient alpha to determine how many raters are necessary for reliable scaling of the subjects. Note that the magnitude of the scalability coefficients is not affected by the number of raters. Alternatively, generalizability theory provides a more extensive selection of methods to investigate reliability (generalizability) of multi-rater tests (see, e.g., Shavelson & Webb, 1991).