Measurement and Prediction in Heterogeneous Populations

Maij-de Meij, A.M.

2008

document version
Publisher's PDF, also known as Version of record

Chapter 6

Conclusion and Discussion

In this chapter, a summary is given of the results of the research described in the different chapters of the thesis. The major conclusions are presented and discussed. Furthermore, suggestions and directions for further research are presented.

6.1 Summary

Latent variable models can be used to describe the response behavior of subjects. Latent trait models are used to account for within-group heterogeneity in the population, whereas latent class models are used to describe between-group differences. As became apparent in this thesis, mixture IRT models may provide a more complete description of test behavior. One can distinguish between quantitative and qualitative differences in the responses of subjects simultaneously. Quantitative differences are conceived as differences in degree, by positioning subjects on a metric scale. Several interpretations of qualitative differences across latent classes have been discussed and analyzed. Also, the value of mixture IRT models for practical applications has been discussed, for the cases of DIF detection and prediction.

In Chapter 2, we studied the concept of self-disclosure. Apart from identifying quantitative differences in the tendency to self-disclose, also identification of qualitatively different self-disclosure patterns was expected. A mixture IRT model with three latent classes was identified. The subjects of the different latent classes varied in their general tendency to self-disclose as well as in their choice to whom they will show self-disclosure. Subjects who respond differently to different categories of people are considered to be selective in their self-disclosure. It was shown that differences in self-disclosure patterns

could be interpreted in terms of differences in selectivity in self-disclosure. Furthermore, extraversion was shown to be associated with the latent trait and latent class variable. However, there was no support for the hypothesis that subjects who are more selective will have lower scores on extraversion. This could be explained by the result that, though contrary to the expectations, subjects with a higher tendency to self-disclose were shown to be the most selective in their self-disclosure. It was demonstrated that it matters whom someone is facing in deciding whether or not to show self-disclosure. We have shown that the mixture IRT modeling framework is a useful tool for identifying differences in kind and degree in responses to personality measurement instruments.

Above, we described the analysis of qualitative differences in the measured attribute. Qualitative differences may also be substantively irrelevant to the measured construct. Differences are defined as methodological artifacts which may reflect item bias. Some items may function differentially across groups of subjects. Generally, DIF detection methods compare the functioning of items across manifest groups (e.g. Camilli & Shepard, 1994; Holland & Wainer, 1993). In Chapter 3, we studied the performance of a manifest DIF detection method in identifying DIF items, using a χ_j^2 statistic (Lord, 1980), and a similar statistic based on the comparison of IRT parameters across latent groups. In a simulation study, we showed that the mixture IRT model performs better in identifying DIF items compared to DIF detection methods using manifest variables only. When there is a high correlation between the manifest variable and the source of bias, both DIF detection methods perform well. However, in situations where the correlation decreases (0.6 or lower) the mixture IRT model is shown to be superior. Furthermore, including the manifest variable as an indicator of the latent class variable has been shown to improve the identification of DIF. An advantage of DIF detection using a latent grouping variable is that one is not restricted to identify DIF associated with a specific manifest variable. The model provides room to detect the true source of bias and can be used even when there is no manifest variable available. Furthermore, when the manifest variable is a valid indicator of the source of bias, it does no harm to include a latent grouping variable.

We have shown that artifacts, like item bias, can very well be analyzed using mixture IRT models. Qualitative differences unrelated to the measured attribute may also reflect different response tendencies. A previous study of the personality scales extraversion and neuroticism by Smit, Kelderman and Van der Flier (2003) showed that the parameters of

a "?" category were not invariant across groups of subjects. In Chapter 4, we extended the study of Smit et al. (2003) by analyzing a larger data set and allowing for the identification of more than two latent classes. A mixture version of the nominal response model with three latent classes was identified as the best fitting model. Response patterns within latent classes demonstrated a differential use of the "?" category. Subjects from one latent class tended to avoid this category, whereas subjects from another latent class did not seem to prefer or avoid any category. Incorporation of covariates in the model provided insight into associations between these covariates and the use of response categories. The latent classes could be characterized by social desirability and ethnic background.

Criterion data, corresponding to the measured attributes, were available for a part of the sample in Chapter 4, which allowed us to study the accuracy of prediction. For the extraversion scale, latent trait scores estimated with a simple IRT model were shown to yield more accurate predictions of the criterion measure than latent trait scores estimated with a mixture IRT model. The neuroticism scale showed more promising results. It was shown that the mixture IRT model could be used to improve the prediction of the criterion, where for two of the latent classes this improvement was significant. It can be concluded that mixture IRT models offer possibilities to improve the prediction of external criteria, though the results are not conclusive.

Subjects are generally assigned to the latent class with the highest probability given their response pattern, after which the corresponding latent trait estimate can be allocated. This procedure is based on the assumption that a subject belongs to one, and only one, latent class (Goodman, 1974). In Chapter 5, we studied an alternative that uses the information of the latent class probabilities for the estimation of latent trait values. This latent trait estimate weighs the class-specific latent trait estimates with the corresponding latent class probabilities. In a simulation study, it was shown that weighted latent trait estimates predict criterion and simulated latent trait values equally well or better than assigned latent trait estimates and latent trait estimates of a simple IRT model. The differences between weighted and assigned latent trait estimates become smaller when subjects are assigned with higher certainty to the latent classes. However, the differences become larger when the class-specific latent trait estimates, before weighting or assignment, vary increasingly. These two opposite effects arise when the differences in discrimination parameters across latent classes become larger. It was

concluded that the weighting procedure performs better, in particular for shorter tests.

Predictions by weighted and assigned latent trait estimates were compared with an empirical data set as well, elaborating on the previous study of two personality scales described in Chapter 4. Before we could proceed with weighting or assignment, the class-specific latent trait estimates were transformed to be on a common scale. For the extraversion scale, again the simple IRT model provided a better prediction of the external criterion measure than the assigned and weighted latent trait estimates, though the differences did not prove to be significant. The weighted latent trait estimates gave a more accurate prediction of the criterion measure compared to assigned latent trait estimates. There were relatively large differences in prediction with weighted and assigned latent trait estimates that could be explained by the smaller variance of individual latent class probabilities and the larger differences in class-specific latent trait estimates. For the neuroticism scale, only the weighted latent trait estimate showed a significantly higher correlation with the criterion compared to the simple IRT model, although the difference in prediction with weighted and assigned latent trait estimates was small. So again, it can be concluded that, depending on the data, the mixture IRT model may improve prediction of external criteria. The results remain mixed, though the weighted latent trait estimate has been shown to offer a possible interesting alternative to assignment of latent trait estimates.

In this thesis, we investigated several applications of the mixture IRT modeling framework. The heart of the models lies in describing within as well as between-group differences, in other words, quantitative and qualitative individual differences. Qualitative differences can be discussed from several perspectives, of which we have focused on situational specificity, item bias and response tendencies. Qualitative differences may be meaningful with respect to the interpretation of an attribute as well as for development of its theoretical framework. When the differences between latent classes are not related to the measured attribute, they may still be meaningful. The influence of response tendencies in personality measurement should not be overlooked. Application of mixture IRT models may provide possibilities to account for these influences while analyzing the latent trait one intends to measure. Promising results were shown for the study of differential functioning of specific items using a latent DIF detection method. No unambiguous results could be obtained from the studies of the mixture IRT modeling framework as a method to improve

estimates of latent trait values and prediction of external criteria.

6.2 Further Research

We have reported on the mixture IRT modeling framework as a procedure to describe quantitative and qualitative individual differences. A general assumption that we made in each of the studies in this thesis, was that the same trait is measured in each of the latent classes. It can be argued that the identification of different item parameters across groups, that is absence of measurement invariance, indicates that different constructs are measured in different latent classes. Indeed, when different constructs are measured, the item parameters can be expected to vary across latent classes. The converse is not necessarily true. When the item parameters vary across latent classes this could also reflect a differential use of the response scale or DIF. In that case, the same attribute is measured across latent classes but in a different way, which was the focus in Chapter 3 through 5. The assumption that within each latent class the same latent trait is measured is important for assigned latent trait estimates to be compared across latent classes, and to compute meaningful weighted latent trait estimates. Of course, the assumption needs to be checked in new studies. Clear criteria need to be specified to determine whether the assumption holds. Furthermore, methods need to be developed to check these criteria.

In the second chapter, where we studied the concept of self-disclosure, the latent trait could be interpreted as describing the tendency to self-disclose. The results exposed latent classes differing in self-disclosure patterns that could be interpreted in terms of selectivity in self-disclosure. Thus, the kind of self-disclosure varied across latent classes. We argued that situational specificity is an important topic in the study of personality, where cross-situational behavior of some subjects may be more consistent than that of others. The modeling framework is also applicable in a broader sense, including the study self-disclosure with respect to people outside the work environment. Furthermore, mixture IRT models may be used to study other personality attributes and their situational dependence as well.

Qualitative differences unrelated to the measured attribute were first of all considered to reflect DIF. We have studied the identification of uniform DIF, where the item difficulty parameters were allowed to vary across two latent or manifest groups. It would be

interesting to extend this to non-uniform DIF, and to compare more than two sets of item parameters simultaneously (see for a multigroup statistic Kim, Cohen, & Park, 1995). Furthermore, a chi-squared statistic (Lord, 1980) was used to identify items of which the parameters differed significantly across groups. There are many other DIF detection methods that have been used to study the differential item functioning across manifest groups, for example the likelihood ratio method (Thissen, Steinberg, & Wainer, 1988; 1993) or area DIF measures (Raju, 1988). These methods can be used to study DIF across latent groups as well. Naturally, empirical studies need to assess the efficiency of the latent DIF detection method for real data sets. As opposed to simulation studies, for real data sets the true source of DIF is unknown, as well as which specific items may display DIF. Therefore, experiments may be conducted, where, for example, two groups of subjects receive different instructions which should result in different responses to specific items (e.g. Kok, Mellenbergh, & Van der Flier, 1985). Then, a mixture IRT model should identify two latent classes, associated with the different instructions. The items that ought to be affected by the different instructions should be identified as displaying DIF across the latent classes.

The simulation study of DIF indicated that even when there is a small correlation between the source of bias and a manifest variable, including this manifest variable as an indicator of the latent class membership facilitates identification of DIF. Incorporation of more than one manifest indicator may be topic for further study. The effect of including more manifest indicators on the number of items identified as displaying DIF should be investigated, or more specifically, its effect on the number of false positives and false negatives. Idealy, taxonomies of personality attributes may be included as indicators of the latent class variable.

In the simulation of Chapter 3 there were a number of biased items to demonstrate the performance of manifest and latent DIF detection methods. However, the number of items that exhibit DIF may affect the identification of DIF items by the two DIF detection methods. A small number of DIF items may limit the latent DIF detection method because of identification problems. When there would be just one item displaying DIF it would be virtually impossible to identify the item using a model with a latent grouping variable and excluding exogenous variables. Including a manifest indicator variable would render the latent class variable superfluous. Yet, there are usually several biased items.

Future research needs to investigate the detection of DIF when there are only a few biased items. This could be examined for different degrees of DIF. In this way, conditions could be made more specific for applying latent as opposed to manifest DIF detection methods.

In Chapter 4, we offered meaningful interpretations of the qualitative individual differences that were unrelated to the measured attribute. Covariates were incorporated in the model to aid in the interpretation of the latent classes. Social desirability and ethnic background were shown to be associated with the differential use of the response scale. Thus, the results may contribute to research into these variables. Of course, the association of the latent classes with many other external variables, like gender, socio-economic status, educational level and so on, may be studied. In addition, measures of personality attributes may be incorporated. Cross-cultural research may benefit from this characteristic of mixture IRT models as well. Recency of immigration, language deficiency, or age of immigration may be incorporated to investigate the connection between latent classes and acculturation.

The use of mixture IRT models for prediction of external criteria needs a great deal of further research. The results of the empirical study were nonconclusive as to what procedure should be followed to make the most accurate prediction of a subjects criterion value. Of course, the criterion may not have been the most valid indicator of the trait to be measured. The study of the empirical data set was based on the fitting of a mixture version of the nominal response model. The model allowed the difficulty and discrimination parameters to vary across items, categories and latent classes. Therefore, it is difficult to isolate factors and determine their specific influence on prediction based on a mixture IRT model. The simulation study in Chapter 5 was based on the two parameter logistic model, where the conditions varied in the extent to which difficulty and discrimination parameters differed across latent classes. It gave us a first view of the conditions that may need to be met to profit from the application of mixture IRT models for the improvement in prediction of external criteria. A simulation study allows for research under specific and controlled situations. Two latent classes of equal size were simulated, and the item parameters were balanced within and between latent classes. The effect of different class sizes and item parameters may be subject for further study. Also, models for polytomous items, that are frequently used in personality assessment, need further investigation.

Recently, Goodman (2007) considered two different procedures for assigning subjects to latent classes. A traditional method that has also been used in this thesis, is the assignment of subjects to the latent class with the highest probability, in other words the modal latent class procedure. The second procedure "... uses random assignments based on the estimated probability distribution of the latent classes corresponding to each of the ... response patterns" (p. 9, Goodman, 2007). The first procedure was demonstrated to minimize the number of incorrect assignments. The second procedure was designed in a way such that the expected proportion of subjects assigned to each of the latent classes would approximate the latent class proportions estimated under the model. To use a strategy of random assignment for estimation of latent trait values is questionable. The weighting procedure we proposed in Chapter 5 seems to be more appropriate, also compared to assignment.

Generally, it is assumed that the same measurement model holds in the different latent classes, and only the item parameters vary across the latent classes. The response behavior of some subjects may be well described by a parsimonious model, like the partial credit model. However, if subjects use the scale differently the nominal response model may be more appropriate to accurately describe their response behavior. As a consequence, a restricted mixture IRT model may be rejected when comparing mixture IRT models that differ in parsimony. Still, the response behavior of subjects of one latent class may be very well described by a simpler IRT model. Models have been developed for scalable and non-scalable subjects, where the same IRT model is specified for scalable subjects (Goodman, 1975; Yamamoto, 1989). A combination of different IRT models in different latent classes has not been studied so far, but may offer new perspectives and opportunities.