Ordered latent class models in nonparametric item response theory
van Onna, Maria Johanna Henrica

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

Introduction

Many people have been tested at least once in their lives. For example, in
school, their abilities in arithmetic, reading, and writing may have been tested.
Personality assessment by means of a test may have taken place during a job
selection procedure. Also, people may have answered to an attitude interview
via the telephone. The use of the test may be of great importance, for example,
when people are selected for a job on the basis of the test result. Given
this importance of a test, its construction should be sound to ensure certain
desirable measurement properties. A lack of quality invalidates the usefulness
of the test, and harms the interests of both the testee and the tester.

Assumptions of item response theory models

The sound construction of tests encompasses not only that the questions and
answers are well formulated, but also that assumptions which are made (either
explicitly or implicitly) in the use of test results are checked on the test data.
The area of statistics which is called item response theory (IRT) deals with
such assumptions for tests. IRT investigates which assumptions are necessary
when using a test in a specific way, and develops statistical methods to check
whether the assumptions are plausible for the group of people which take the
test.

The first central assumption in IRT is that the test should measure only
one ability, personality trait or attitude. This is called the assumption of uni-
dimensionality (UD). The unidimensional ability, personality trait or attitude, which is measured by a test, is generally denoted as the latent trait in IRT. Only if the test is unidimensional, the test result can be used to make statements about the trait level of each person. For example, a reading test may only result in valid reading grades for each person, when no other latent trait (like knowledge of the topic of the text) influences the probabilities of responding correctly to the test items. Multidimensional IRT models also exist, but the interpretation of their results is more complicated since multidimensional models do not result in one unique ordering of subjects.

The second central assumption is the assumption of local independence (LI). This assumption means that the responses to the items are independent of each other given a fixed location on the latent trait. That is, association between two items can only be attributed to variation on the latent trait between people. LI is strongly related to the assumption of unidimensionality, in that no other ability or property of the people who take the test influences the response probabilities.

An important aspect of tests which is crucial in IRT, is the relation between the latent trait which is measured by the test, and the response probabilities to the items. This relation is expressed by means of the item response function (IRF). The probability of responding positively to an item $j$ is expressed as a function of the latent trait $\theta$: $P(X_j = 1|\theta) = f_j(\theta)$, where $f_j(\theta)$ is the IRF (Figure 1.1). The distinction between parametric IRT and nonparametric IRT is in the formulation of the IRF. In parametric IRT, each $f_j(\theta)$ is a continuous function such as the logistic function or the normal ogive, with parameters indicating the location, the slope and the asymptotic values. In nonparametric IRT, only order restrictions are imposed on IRFs. The main nonparametric assumption on the shape of the IRF is that it is non-decreasing. This is called the assumption of monotonicity (M).

A second assumption on the IRFs, which is common in nonparametric IRT, is the assumption of non-intersection (NI). It states that the IRFs of different items do not intersect. The difficulty of an item can be expressed as the population proportion of incorrect responses, $1 - P(X_j = 1)$, because an item is difficult if many people give an incorrect answer. NI implies that the item
Measurement errors may occur in the ordering of people. The number of measurement errors is high, if the items hardly discriminate between low-ability and high-ability persons, that is, if the IRFs are almost flat. Mokken (1971) developed a scaling coefficient $H$, which indicates how well the ordering of the persons on the latent trait is established. Under the null-model of (globally) independent items, that is, if the items would not be related to each other because they do not measure the same latent trait, the value of $H$ is zero. If the ordering of the persons is perfect, $H = 1$. In this case, the covariances between the items are as large as possible given the marginal frequencies. If $0.3 \leq H < 0.4$, the test is a weak scale. If $0.4 \leq H < 0.5$, the scale is medium, and $H \geq 0.5$ indicates a strong scale (Mokken, 1971, p. 185).

Another concept which deals with the amount of error in test results is reliability. In classical test theory, Cronbach’s coefficient $\alpha$ is used as a lower bound estimate of this reliability. It equals zero when the items are not correlated to each other. Its maximal value is one. Although both $H$ and $\alpha$ may be interpreted as coefficients of internal consistency, they are not identical. Their behavior has been compared extensively, for example by Molenaar and Sijtsma (1984). In nonparametric IRT, another method was developed by Sijtsma and Molenaar (1987) to estimate the reliability of a test. This method requires that the assumptions of UD, LI, M and NI are fulfilled.

So far, we only discussed dichotomous items, which have two response options like ‘correct’ and ‘incorrect’, or ‘yes’ and ‘no’. A polytomous item has more than two response options. This is for example the case in attitude items, where one can indicate to which degree one agrees with a statement (‘not at all’, ‘a little’, ‘fairly’, ‘(almost) completely’). At some places in this thesis, polytomous items are studied. The use of $X_+$ to order people depends on an additional property of the test, known as stochastic ordering of the latent trait by the total score (SOL-$X_+$). SOL-$X_+$ means that $P(\theta > c|X_+)$ is increasing in $X_+$, given a fixed value for $c$. It is guaranteed for dichotomous items, when UD, LI and M hold. However, for polytomous items some violations of SOL-$X_+$ may occur. A violation of SOL-$X_+$ usually occurs in the extremes of the distribution of the total score $X_+$. However, not many people have these extreme scores, and consequently a violation of SOL-$X_+$ does not affect the
ordering of most people (Van der Ark, in press).

**Model testing in NIRT**

The assumptions of UD, LI and M may be checked specifically, in order to evaluate whether they are justified. For example, the MSP computer program (Molenar & Sijtsma, 2000) offers the possibility to check assumption M by means of the regression of the response probability on the restscore, which is the total score $X_+$ minus the score on the item of which the response probabilities are investigated. The statistic $\text{crit}_j$ is a weighted sum of the sizes and significance levels of all decreases in the estimated IRF of item $j$. If it is at least the suggested critical value of 80, it can be concluded that M is violated for item $j$. The plot of the IRF can be made in a smoother way, by using splines or kernel smoothing techniques (Ramsay, 1991b, 1991a). A visual inspection then has to decide whether (local) decreases in the IRF occur.

In Chapter 4 of this thesis, several nonparametric techniques for investigating the assumption of UD are considered. The search option of the MSP computer program and the program detect (Zhang & Stout, 1998) both cluster items into unidimensional sets. When more than one cluster, each with more than one item, is formed, the UD is violated. Two other NIRT methods are based on ordered latent class models, which will be discussed briefly below. If the fit of a 1-factor latent class factor model is as good as the fit of a 2-factor model, it is concluded that unidimensionality holds. Similarly, UD is accepted when posterior predictive checks indicate that a unidimensional ordered latent class model fits the data.

**Ordered latent class models**

In this thesis, ordered latent class models are studied within the context of NIRT. In ordered latent class models, it is assumed that only a restricted number of trait levels exist. For example, two discrete skill classes may be discerned, the ‘non-masters’ and the ‘masters’. On a depression scale, one might consider ordered classes ‘not depressed’, ‘slightly depressed’ and ‘severely depressed’. A dimension indicating the attitude towards abortion may have the
levels ‘always pro abortion’, ‘in most situations pro abortion’, ‘in most situations against abortion’, and ‘always against abortion’. People with the same trait level can be clustered in homogeneous groups. These groups are called latent classes, which may be considered as a typology of people. The latent classes are ordered, because the trait levels can be ordered from ‘low’ to ‘high’.

The response probability on an item is assumed to be the same for each person in a latent class. The IRF can now be plotted as a discrete function of the class levels (see Figure 1.1). One reason to consider an ordered latent class model as an NIRT model is that the discontinuous function can be used to approximate most continuous functions quite well, if the number of latent classes is large enough. Another reason is that a test with a small number of items has only a restricted range of values of the total score $X_\pm$. The ordering of people who take the short test then is not really more refined than the ordering by means of ordered latent classes.

The advantage of ordered latent class models, in comparison to continuous latent trait NIRT models, is that they allow for a full parameterization of the IRFs, without losing the generality of IRFs which are only order-restricted. By fitting an ordered latent class model, the opportunity exists to estimate and check a nonparametric IRT model with statistical methods. The number of parameters, however, is large. For $Q$ latent classes and $J$ dichotomous items, $Q - 1$ class weights and $Q \times J$ conditional response probabilities have to be estimated. For polytomous items each with $k$ response categories, even $Q \times J \times (k - 1)$ conditional response probabilities have to be estimated. This huge parameter space is narrowed down by the order restrictions on the conditional response probabilities, imposed by the assumptions of M and NI. This limiting effect of order restrictions is both an advantage and a disadvantage for model estimation and model checking. The advantage is a more directed search for parameter estimates. The disadvantage is that standard maximum likelihood estimation and model checking procedures do not work. However, a complicated Markov chain Monte Carlo method, which is a Bayesian approach, gives the opportunity to estimate and check ordered latent class models.
Contents of this thesis

In Chapter 2, Bayesian estimation and model checking of ordered latent class models is described. In classical statistical theory, a point estimate is made of each parameter based on the data. It reflects the value of a parameter which fits best to the data, given some criterion like maximum likelihood or least squares. In Bayesian estimation, the information on each parameter is expressed in a distribution. An initial guess on the possible values of a parameter is expressed in a prior distribution. It is adapted after processing the data, resulting in a posterior distribution. If the initial guess is vague, resulting in uninformative priors, the mode of a posterior distribution will equal a maximum likelihood estimate. That is, although the conceptual differences between a classical and a Bayesian approach may be enormous, the practical differences may be smaller. In ordered latent class models, the order restrictions on the parameters are quite easily implemented in the Bayesian estimation algorithm of the posterior distribution. A classical point estimate may be difficult to obtain under the order restrictions. The Bayesian iterative algorithm is based on the Gibbs sampler (see e.g. Casella & George, 1992), which is a Markov chain Monte Carlo (MCMC) algorithm.

The Bayesian estimation algorithm is accompanied by Bayesian model checking tools such as posterior predictive checks and Bayes factors (Chapter 2). The posterior predictive checks can be used to evaluate models globally, but also to evaluate the fit of one item. In Chapter 3, the global and specific misfit of NIRT models is investigated by means of two methods: the Bayesian ordered latent class approach, and Mokken scale analysis. In Chapter 4, posterior predictive checks of the ordered latent class model are contrasted with three other ways of detecting violations of UD, the detect algorithm and statistic, the MSP search algorithm, and latent class factor analysis.

The adequacy of a test for ordering persons may be evaluated by means of coefficient $H$. This coefficient is easy to estimate given a sample. However, the sampling distribution of $H$ is only known asymptotically, and under certain restrictions. In Chapter 5, the sampling distribution of the scaling coefficient $H$ is discussed. Several estimation methods are compared, among which estimation methods based on a fitted ordered latent class model. Both classical
and Bayesian methods are used. The other methods which are discussed are
a naive bootstrap method and the asymptotic variance as derived by Mokken
(1971, p. 166).

Several chapters of this thesis have been published or are submitted for
publication. Chapter 2 appeared as Van Onna (2002), Chapter 3 as Van Onna
(2003). Chapters 4 and 5 have been submitted. This may explain some overlap
between the chapters. All references in the papers have been collected into
one common list.