# Evaluating model assumptions in item response theory

## Het evalueren van modelassumpties in item response theorie

Jesper Tijmstra

ISBN: 978-90-8891-721-9

| | |
|---|---|
| Cover design: | Jesper Tijmstra & Proefschriftmaken.nl \|\| Uitgeverij BOXPress |
| Lay Out by: | Jesper Tijmstra & Proefschriftmaken.nl \|\| Uitgeverij BOXPress |
| Printed by: | Proefschriftmaken.nl \|\| Uitgeverij BOXPress |
| Published by: | Uitgeverij BOXPress, 's-Hertogenbosch |

# Evaluating model assumptions in item response theory

## Het evalueren van modelassumpties in item response theorie

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht op gezag van
de rector magnificus, prof. dr. G. J. van der Zwaan,
ingevolge het besluit van college voor promoties
in het openbaar te verdedigen op vrijdag 15 november 2013
des middags te 12:45

door

**Jesper Tijmstra**

geboren op 8 juni 1985
te Amersfoort

| | |
|---|---|
| Promotoren: | Prof. dr. P. G. M. van der Heijden<br>Prof. dr. K. Sijtsma |
| Co-promotor: | Dr. D. J. Hessen |

# Contents

# Chapter 1

# Introduction

## 1.1 Measuring Latent Attributes

Measurement plays a central role in most of the empirical sciences. Ranging from the tiniest particles to galaxies at the outermost reaches of space, scientists want to measure every phenomenon they encounter. The social and behavioral sciences are not very different in this respect: From an individual's level of intelligence, extraversion or depression, to the attitudes of whole societies towards issues of gender equality and euthanasia, most scientists in this field are concerned with the measurement of the complex constructs that feature in their elaborate theories. What makes measuring these constructs especially difficult is that almost none of the attributes that these theories deal with can be observed directly. There is simply no direct way of measuring someone's intelligence in the way that one is able to measure someone's weight or eye color. That is, intelligence is a latent property for which there is no direct form of measurement available. This makes the question of how we can successfully measure these elusive constructs one of the most important and challenging questions facing the social and behavioral sciences.

Although we do not have a direct way of measuring constructs such as intelligence and extraversion, we can still make inferences about these

latent constructs by considering the way in which they relate to observable phenomena. Extraverted people tend to seek out social activities more than introverted people, so by finding out to what extent someone engages in this kind of activities we are able to obtain information that may be relevant for the assessment of that person's level of extraversion. Likewise, intelligent people tend to do intelligent things, so by putting them in a situation where they can display intelligent behavior – having them answer questions on an intelligence test – we can attempt to gain insight into their level of intelligence.

Still, translating such observational information into an accurate assessment of the latent property that we are interested in is a complex affair. For example, simply noting that someone answered 7 out of 10 questions correctly on an intelligence test does not tell us much about that person's intelligence if we do not know how difficult these items were. To make the step from information at the manifest level – the level of observations – to inferences at the latent level, statistical models are needed. Fortunately, psychometricians and statisticians have developed a wide range of models that aim to facilitate this difficult inferential step.

With the emergence and growth of item response theory (IRT) since the 1950s, a plethora of measurement models have been developed that provide powerful tools for the measurement of latent properties based on discrete response data. These IRT models make it possible to estimate a person's ability or value on a latent trait, and to critically evaluate the properties of the items that are included in the scale or test that is used to measure this construct. One of the main attractive features of IRT is that it places both persons and items on the same latent scale. In the same way that a person may have a high or low ability or trait level, an item can have a high or low difficulty and can be located on the same scale. Having a common scale for persons and items makes it possible to assess the amount of information that items provide for specific ranges of the latent variable, and also to adaptively match items to the estimated ability level of the person taking the test, as in adaptive testing (Van der Linden & Glas, 2010). An additional advantage of IRT models is that they ensure that the expected ordering of persons does not depend on the specific set

of items that is used, which makes the use of item banks possible. For a more extensive overview of IRT, the reader is referred to Lord and Novick (1968), Lord (1980), Hambleton and Swaminathan (1985), and Van der Linden and Hambleton (1997). In short, IRT provides powerful tools for the assessment of both persons and items, and because of these benefits it has become the dominant approach in psychometrics to measuring latent constructs.

## 1.2 Assumptions in Item Response Theory

As with any statistical model, IRT models are defined by their assumptions. In particular, most of the assumptions in IRT concern the item response function (IRF), which describes the relationship between the latent variable (or variables) and the probabilities of the responses that are obtained on the items on the scale or test.

Let $\boldsymbol{\theta}$ denote a vector containing the set of latent variables relevant to responding to the items, and let $X_i$ be the random variable for the score on an item $i$ (with realization $x_i$). The IRF of a dichotomous item $i$ (e.g, an item that is answered either correctly or incorrectly) corresponds to

$$P(X_i = 1|\boldsymbol{\theta}). \tag{1.1}$$

Thus, the IRF describes the probability of obtaining a positive response on an item $i$ (e.g., successfully answering a math exam question) conditional on the latent variable(s) (e.g., general math ability or a set of specific math skills). The restrictions that are placed upon the IRF define the specific IRT model. Whether it is plausible to assume that these restrictions hold for a particular application of a test in a population is an important question, and this dissertation focusses on developing methods that can be used to assess assumptions concerning the IRFs.

### 1.2.1 Local Independence

An assumption that is shared by almost all IRT models is local independence (also known as conditional independence), which states that after

taking the latent variables $\boldsymbol{\theta}$ into account, the responses to the items are independent of each other. If we denote the vector containing the scores on each of the $k$ items in the test by $\boldsymbol{X}$ (with realization $\boldsymbol{x}$), the assumption of local independence corresponds to

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{k} P(X_i = x_i|\boldsymbol{\theta}). \tag{1.2}$$

Items in a test are usually developed to measure one or more common latent variables. Because of this, there generally is a positive association between the responses to these items. This positive association is desirable, since it indicates that there are factors that are shared by the items, and the hope is that these factors correspond to the constructs that the scale was designed to measure. The idea behind the assumption of local independence is that once we have taken these latent variable(s) into account, the responses to the items no longer depend on each other: They are independent conditional on the latent variable(s). That is, while the items are globally dependent (i.e., over the full range of the latent variable(s)), they are locally independent (i.e., if one controls for the latent variable(s)).

Local independence may be violated when two items on a test are very similar. For example, if in a math test elementary school students have to solve both '$42 * 27 =$ ?' and '$27 * 42 =$ ?', many students may realize that the answer should be the same. In that case, a student's probability of providing a positive response to the second item may depend on his response to the first item, even after we take general multiplication ability into account.

A variety of tests for local independence have been developed (see e.g. Lord, 1980; Holland & Rosenbaum, 1986; Stout, 1987). While this dissertation does not focus on the evaluation of local independence, the assumption of local independence is crucial for the derivation of many of the observable consequences that are used in the testing procedures for IRF assumptions that are proposed in this dissertation. This means that if these tests indicate that an assumption is violated in a particular data set, this conclusion can only be drawn under the assumption that local independence holds.

Thus, the application of the tests for IRF assumptions proposed in this dissertation should always occur in tandem with tests that evaluate local independence. This way, the user can obtain information about which particular IRT assumption may be violated.

### 1.2.2 Unidimensionality

The assumption of unidimensionality is strongly related to the assumption of local independence. Unidimensionality tells us that the dependence between the item scores can be fully explained by a single latent variable. That is, if we would know someone's value on the latent variable, there are no other relevant influences on the probability of observing a specific item score. If we denote this single latent variable by $\theta$, then for any set of additional variables $\boldsymbol{Z}$ with realization $\boldsymbol{z}$ we obtain

$$P(\boldsymbol{X} = \boldsymbol{x}|\theta) = P(\boldsymbol{X} = \boldsymbol{x}|\theta, \boldsymbol{Z} = \boldsymbol{z}), \text{for all } \boldsymbol{z}. \tag{1.3}$$

When unidimensionality holds, local independence also holds (Hambleton & Swaminathan, 1985). That is, Equation 1.3 tells us that beyond $\theta$ there are no other influences on $P(\boldsymbol{X} = \boldsymbol{x})$. If there would be dependencies between the item scores that are not explained by taking $\theta$ into account, we could always introduce another variable that explains this dependency. However, Equation 1.3 informs us that $\theta$ is the only variable that is relevant for explaining the dependencies in the data, and hence local independence follows from unidimensionality.

In many applications of IRT, the assumption of unidimensionality is attractive. Many scales or tests are developed to measure a single trait or ability, and if unidimensionality holds we know that the item scores can be explained by means of one latent variable. This does not prove that the latent variable correctly represents the attribute that the researcher hopes to measure, but based on content analysis of the items in the scale and by relating the latent variable to other attributes it can be made plausible that the latent variable corresponds to the construct he aimed to measure.

Although unidimensionality often is an attractive assumption, it may be violated in many situations. In real life, the answer to an item will often

depend on more than one attribute. Many people are familiar with extensively worded questions in math exams, where one has to solve a mathematical problem that is embedded in a story. For such an item, mathematical ability may not be the only factor that influences performance, and we may have to take reading ability into account as well. If this is the case, unidimensional IRT models cannot be used.

If unidimensionality is violated, this means that attributes other than the ones that we try to measure influence the responses to the items, which could result in unfair or biased conclusions about persons. For example, it could be that for a specific item in an intelligence test men structurally outperform women. On closer inspection, it might turn out that the question revolves around an example from sports, and that in order to correctly answer the question one has to be familiar with the rules of that sport. In such a situation, we may be dealing with differential item functioning (DIF), since the performance on the item differs for men and women. If this is the case, then whether one is able to correctly answer the question does not only depend on intelligence but also on gender, and this violation of unidimensionality biases the test against females.

In the example of the worded math questions it is plausible that correctly answering these questions depends on at least two abilities: math skill and comprehensive reading skill. In such a situation, unidimensional IRT models will not suffice, and one may have to use multidimensional IRT. The field of multidimensional IRT generalizes the IRT framework to situations where multiple latent variables have to be taken into account (see e.g. Reckase, 2009). A variety of tests have been developed that can help determine whether multidimensional models need to be applied, or whether a unidimensional IRT model may suffice (Rosenbaum, 1984; Hattie, 1985; Hambleton & Rovinelli, 1986; Gessaroli & De Champlain, 1996; Roussos, Stout & Marden, 1998). Because most scales and tests are developed with the aim of measuring a single ability or trait, the main focus of this dissertation will be on the evaluation of model assumptions for unidimensional IRT models. Additional research should enable the generalization of most of the methods that are proposed in this dissertation to multidimensional IRT applications.

### 1.2.3 Latent Monotonicity

While unidimensionality specifies that there is one latent variable that governs the probability of observing a certain score on an item, it does not tell us how the probability of observing that score relates to the latent variable. In order for the items to provide us with information about someone's value on the latent variable and hence to obtain a measurement model (Suppes & Zanotti, 1981; Holland & Rosenbaum, 1986), a further constraint needs to be imposed on the shape of the IRF. For this purpose, the assumption that is most commonly made is latent monotonicity. In the context of dichotomous IRT, latent monotonicity means that the probability of observing a positive or a correct response to the item increases as the latent variable increases. Thus, for an item $i$ latent monotonicity corresponds to

$$P(X_i = 1|\Theta = \theta_a) \leq P(X_i = 1|\Theta = \theta_b), \text{for all } \theta_a < \theta_b. \tag{1.4}$$

Equation 1.4 informs us that latent monotonicity also allows the probability of obtaining a positive score to be constant across some range of the latent variable. While strictly speaking Equation 1.4 also allows the IRF to be constant across the entire range of the latent variable, such items would in practice be removed from the scale since they do not provide information about the latent variable.

The main motivation behind assuming latent monotonicity is that the assumption captures the idea that the items measure the latent variable (Junker & Sijtsma, 2000). If latent monotonicity is violated, at least for some range of the latent variable the probability of obtaining a positive response decreases as the latent variable increases. For example, for an exam item nonmonotonicity means that high-ability students may have a *lower* probability of providing the correct answer than students with a lower ability. Intuitively, this tells us that the item does not fully succeed in measuring ability.

The combination of the assumptions of local independence, unidimensionality and latent monotonicity provides us with one of the most general IRT models (Junker & Sijtsma, 2000). This model is known as the monotone homogeneity model (Mokken, 1971), but it has also been studied under

a variety of different names (see e.g. Holland & Rosenbaum, 1986; Junker, 1993; Ellis & Junker, 1997; Van der Linden & Hambleton, 1997). The model has a set of attractive properties, most notably the monotone likelihood ratio of the unweighted sumscore of the items conditional on the latent variable (Grayson, 1988; Huynh, 1994). This property implies that obtaining a high sumscore on the test becomes more and more likely as the value on the latent variable increases. From this monotone likelihood property it follows that the unweighted sumscore – also known as the total score – stochastically orders persons on $\theta$, regardless of the subset of items from the test that is used. That is, one could pick any subset of items for which the monotone homogeneity model holds, and whenever one person obtains a higher total score than another person, the first person is more likely to have a higher value on the latent variable than the second person. Thus, the monotone homogeneity model ensures that the total score constitutes an ordinal scale. This makes it an attractive measurement model for many applications where the goal is to order persons by ability, such as cases where the best 10 candidates or students should be selected.

Several authors have proposed tests for latent monotonicity (Mokken, 1971; Rosenbaum, 1984; Ramsay, 1991; Abrahamowicz & Ramsay, 1992; Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002; Scheiblechner, 2003; Karabatsos & Sheu, 2004). The main difficulty in evaluating latent monotonicity is that the latent variable is unobservable, and that evaluating Equation 1.4 cannot be done in a direct way. Rather, any test of latent monotonicity has to make use of the information available at the manifest level, that is, the observed item scores. Latent monotonicity implies a variety of observable properties that have to hold at the manifest level. The existing tests evaluate these observable consequences, and use information about the observable consequences to make inferences about latent monotonicity.

### 1.2.4 Invariant Item Ordering

The assumption of latent monotonicity tells us that we are able to stochastically order persons regardless of the specific subset of items that we use

from the set of items for which the monotone homogeneity model holds. A similar assumption can be made concerning the ordering of the items. That is, similar to the way that the items order persons under latent monotonicity, one can assume that any subset of persons orders the items based on their expected value. Sijtsma and Junker (1996) call this property an invariant item ordering. If we number the $k$ items on a scale based on increasing overall difficulty, the assumption of invariant item ordering for dichotomous items amounts to

$$P(X_1 = 1|\theta) \leq P(X_2 = 1|\theta) \leq \ldots \leq P(X_k = 1|\theta), \text{for all } \theta. \tag{1.5}$$

Thus, the assumption of invariant item ordering states that none of the IRFs intersect with any of the $k-1$ other IRFs. This means that whenever one item is more difficult than another item for a given value on the latent variable, there is no possible value on the latent variable where this order is reversed. However, since Equation 1.5 does not specify strict inequalities, it does allow for the possibility of having fully or partially coinciding IRFs, as long as they do not intersect.

By adding the assumption of invariant item ordering to the previous three assumptions, a special case of the monotone homogeneity model is obtained, which is the double monotonicity model (Mokken, 1971; Sijtsma & Junker, 1996). The measurement-theoretical appeal of this model is that it allows for the ordering of items as well as persons. Having an invariant item ordering can be useful in practice, for example in the context of person fit and DIF-analysis, or in applications where starting and stopping rules are used (Sijtsma & Junker, 1996). Furthermore, an invariant item ordering is sometimes implied by substantive theory about the content of the items (Sijtsma & Molenaar, 2002). In general, having an invariant item ordering facilitates the interpretation of the test data.

If invariant item ordering is violated, an analysis of the content of the items may be called for to find out why an item is relatively more difficult for a part of the range of the latent variable. A variety of tests for invariant item ordering have been proposed (Mokken, 1971; Rosenbaum, 1987a; 1987b; Sijtsma & Meijer, 1992; Scheiblechner, 2003; Ligtvoet, 2010; see also Sijtsma & Junker, 1996).

### 1.2.5 Parametric Form of the Item Response Function

Both latent monotonicity and invariant item ordering only specify order restrictions for the IRFs, leaving the precise shape of the IRFs open. As a result, the monotone homogeneity model may tell us that one person has a higher ability than another person, but the model cannot inform us precisely how large the difference in ability is. Likewise, the double monotonicity model may inform us that one item is uniformly more difficult than another item, but the model does not inform us about the distance between the items on the scale. Depending on the particular application of the model that one has in mind, this ordinal level of measurement may or may not suffice.

If it is not only important that persons are ordered but also that we have an accurate estimate of their ability or their value on a trait, one has to make additional assumptions about the shape of the IRF by positing the IRF to have a specific parametric form. By adding the assumption of a parametric form of the IRF, we move away from the so-called nonparametric IRT models (Sijtsma & Molenaar, 2002) and obtain a parametric IRT model. Using parametric IRT models, it is possible to obtain an interval level of measurement, and quantitative inferences can be made about a person's value on the latent variable.

One of the simplest and most commonly used parametric IRT models is the Rasch model (Rasch, 1960; Lord & Novick, 1968, p. 402), which states that each person and each item can be characterized by a single parameter value. In the case of persons, this parameter $\theta$ corresponds to their ability level. The items are characterized by their difficulty, here denoted by $\beta$. Under the Rasch model, the probability of person $j$ giving a positive response to item $i$ is obtained through

$$P(X_i = 1|\theta_j) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}. \tag{1.6}$$

Many extensions of the Rasch model have been proposed (see e.g. Lord & Novick, 1968; Hessen, 2006; Von Davier & Carstensen, 2007), generalizing Equation 1.6 by including more parameters to allow for a wider range

of possible IRFs. Since the Rasch model may not be realistic for every application, these extensions allow for greater flexibility and may be useful when a Rasch model does not sufficiently capture the structure of the data. However, regardless of the specific form of the IRF, these models all share the assumption that the IRF of each of the items has the same parametric form. This assumption is usually evaluated by taking the model fit into account, and by comparing the model with other parametric models.

## 1.3 Aim and Outline of the Dissertation

This dissertation deals with the evaluation of latent monotonicity and invariant item ordering. While there are already a variety of tests available for these assumptions, there is still no ideal way of evaluating them. The procedures proposed in this dissertation aim to overcome some of the weaknesses that the existing methods share, and aim to provide powerful new tests for these two assumptions.

The procedures proposed in this dissertation do not require the IRF to have a particular parametric form, and hence remain neutral with regard to the question whether a parametric or nonparametric IRT model should be applied to the data. The motivation for this approach is the idea that tests for latent monotonicity and invariant item ordering should evaluate these properties without already assuming that a specific parametric IRT model best describes the structure in the data. If the tests depend on a particular parametric specification of the IRF, then it may for example remain unclear whether a significant test result indicates that latent monotonicity may be violated or whether the parametric form of the IRF has been misspecified. In such a situation, it is unclear whether we should conclude that the application of any IRT model that assumes latent monotonicity is unwarranted, or whether we should reject the particular IRT model that was used and apply a more general IRT model. For this reason, the procedures proposed in this dissertation do not provide tests of model fit for a particular IRT model, but rather focus on the evaluation of specific properties of the IRFs (i.e., whether they are monotone and nonintersecting). This way, the pro-

posed procedures can help users determine whether specific assumptions about the IRFs can be made, and hence the proposed procedures can be used before the user chooses a specific IRT model.

The tests that are proposed in this dissertation can thus be applied both in the context of nonparametric IRT and parametric IRT, since the model assumptions that are evaluated here are shared by both approaches. For example, if latent monotonicity is violated, both the Rasch model and the monotone homogeneity model are invalid. Thus, the procedures proposed in this dissertation have a broad range of applications, and can be used to determine whether the application of a parametric or nonparametric IRT model may be reasonable.

Chapter 2 proposes a frequentist test for latent monotonicity, making use of the order-constrained statistical inference framework. With this procedure, it is possible to test whether the data indicate that manifest monotonicity – which is an observable consequence of latent monotonicity – is violated. This test is able to detect global as well as local violations of latent monotonicity, and it is shown to have high power to detect these types of violations under a range of reasonable conditions. The performance of the procedure is evaluated using a simulation study, and the application is illustrated using an example from developmental psychology.

In Chapter 3, an alternative procedure to evaluating manifest monotonicity is proposed. This method makes use of the Bayes factor to assess the amount of support that manifest monotonicity receives over some alternatives. A two-step procedure is proposed to determine whether manifest monotonicity is supported by the data. Using Gibbs samplers, it is first determined whether the hypothesis that manifest monotonicity holds receives sufficient support over its negation. If the amount of support is satisfactory, one can proceed to contrast manifest monotonicity with relevant alternative hypotheses. This way, one is able to critically assess the possibility that there are local violations of manifest monotonicity, even if the IRF is predominantly monotone.

Chapter 4 focusses on the assumption of invariant item ordering. It is proven that an observable consequence of invariant item ordering can be derived without assuming a parametric shape of the IRF, and this property

can be used to test for invariant item ordering under very general conditions. A nonparametric measure of invariant item ordering is proposed, and a Bayesian procedure that can be used to test invariant ordering is discussed. The application of this procedure is illustrated using a real-data example from developmental psychology.

Chapter 5 deals with the philosophical question of how one can determine whether a model assumption is plausible. It is argued that if we want to place trust in inferences that are made using a particular statistical model, we have to be sufficiently convinced that the model's assumptions hold for the application at hand. However, it is argued that this decision cannot be based on the data alone, since the plausibility of the model assumption also depends on other considerations. The plausibility of the model assumption before observing the data may heavily influence the plausibility of the assumption after having observed the data. If we are to claim that we consider our model assumptions to be sufficiently justified and that we can trust inferences that are made using the model, it is necessary to take this prior plausibility of the model assumptions into account. This means that null hypothesis-based tests alone may not always provide sufficient information about the model assumption to make a decision whether we can use the model for the application envisaged.

The final chapter concludes with a short summary of the main findings of the dissertation, and some remarks about their consequences. Suggestions for further research are discussed.

# Chapter 2

# Testing Manifest Monotonicity Using Order-Constrained Statistical Inference

Most dichotomous item response models share the assumption of latent monotonicity, which states that the probability of a positive response to an item is a nondecreasing function of a latent variable intended to be measured. Latent monotonicity cannot be evaluated directly, but it implies manifest monotonicity across a variety of observed scores, such as the restscore, a single item score, and in some cases the total score. In this study, we show that manifest monotonicity can be tested by means of the order-constrained statistical inference framework. We propose a procedure that uses this framework to determine whether manifest monotonicity should be rejected for specific items. This approach provides a likelihood ratio test for which the $p$-value can be approximated through simulation.

This chapter has been published as: Tijmstra, J., Hessen, D. J., Van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statisticial inference, *Psychometrika, 78,* 83–97.

A simulation study is presented that evaluates the Type I error rate and power of the test, and the procedure is applied to empirical data.

## 2.1 Introduction

A very general dichotomous item response theory (IRT) model for the ordinal measurement of a continuous latent variable is Mokken's nonparametric monotone homogeneity model (Mokken, 1971). This model is characterized by three assumptions: unidimensionality, local independence and latent monotonicity. The first two assumptions state that all possible associations between the items can be explained by a single latent variable representing a trait or an ability. Latent monotonicity specifies that the probability of observing a positive response to an item increases monotonely as the latent variable increases. Special cases of the model are the well-known one-, two- and three-parameter logistic models (Rasch, 1960; Birnbaum, 1968) and Mokken's non-parametric double monotonicity model (Mokken, 1971).

From the assumptions of the monotone homogeneity model it follows that the total score – the unweighted sum of the item scores – has the property of monotone likelihood ratio with respect to the latent variable (Grayson, 1988; Huynh, 1994; Ünlü, 2008), which ensures that the total score stochastically orders persons on the latent variable (Hemker, Sijtsma, Molenaar & Junker, 1997). As such, the model captures the idea that the items measure the latent variable on an ordinal level (Junker & Sijtsma, 2000). Thus, when the purpose of measurement is to order individuals on the latent variable, it is important to have a means at one's disposal to check whether the model assumptions hold. In this study, a means is provided to check the assumption of latent monotonicity using the order-constrained statistical inference framework (see, e.g., Silvapulle & Sen, 2005).

Let $X_1, \ldots, X_k$ denote the scores on $k$ dichotomous items, let $\theta$ denote the continuous latent variable, and let $\theta_a$ and $\theta_b$ denote two values of the latent variable. The assumption of latent monotonicity can be formulated as

$$P(X_i = 1|\theta_a) \leq P(X_i = 1|\theta_b), \quad \text{for all } \theta_a < \theta_b \text{ and all } i, \tag{2.1}$$

or equivalently, as nondecreasingness of $P(X_i = 1|\theta)$ in $\theta$, for all $i$.

Since $\theta$ cannot be observed directly, Equation 2.1 cannot be evaluated directly. However, under certain circumstances latent monotonicity can be evaluated indirectly by considering monotonicity over a manifest score,

$$Y = \sum_{i=1}^{k} c_i X_i, \tag{2.2}$$

where $c_1, \ldots, c_k$ are item coefficients that are chosen by the researcher, and $c_i \in \{0, 1\}$ for all $i$. Different choices for these coefficients define different manifest scores. For example, one could consider manifest monotonicity for a particular item $j \neq i$ (Mokken, 1971) by choosing $c_j = 1$ and setting all other coefficients to 0. Alternatively, one could use the unweighted restscore (Junker & Sijtsma, 2000), which for an item $i$ is obtained by setting $c_i = 0$, and $c_j = 1$ for all $j \neq i$. Another option would be to focus on monotonicity over the unweighted total score (Hessen, 2005), by letting $c_j = 1$, for $j = 1, \ldots, k$.

For every set of coefficients that can be selected, the highest possible manifest score that can be attained – denoted by $h$ – corresponds to

$$h = \sum_{i=1}^{k} c_i.$$

There are, thus, $h + 1$ possible realizations or 'levels' of $Y$, and this number depends on which coefficients are used in Equation 2.2. Using this formulation, manifest monotonicity over $Y$ corresponds to

$$P(X_i = 1|Y = 0) \leq P(X_i = 1|Y = 1) \leq \ldots \leq P(X_i = 1|Y = h), \quad \text{for all } i. \tag{2.3}$$

As long as one excludes the item under consideration from the manifest score that is selected, latent monotonicity and local independence together imply Equation 2.3 (Rosenbaum, 1987b). Thus, a statistical test that evaluates this property for every item can be used as a test for the monotone homogeneity model. If one includes the item under consideration in the

manifest score, one constructs a test for the Rasch model instead, since monotonicity across the total score necessarily holds for the Rasch model (Rasch, 1960; Hessen, 2005) but can be violated under less restrictive models such as the two-parameter logistic model (Birnbaum, 1968; Junker & Sijtsma, 2000). However, given the multitude of tests for the Rasch model already available (see, e.g., Glas & Verhelst, 1995), the main value of the proposed test lies in evaluating the monotone homogeneity model, and for this purpose the item under consideration should be excluded from the manifest score (Junker & Sijtsma, 2000).

This article provides a test for manifest monotonicity for individual items. The current approach has added value over existing approaches to evaluating latent monotonicity, since these other approaches either require the evaluation of many partial results due to the partitioning of the test into separate sets of items (see, e.g., Rosenbaum, 1984) or involve test statistics of which the distribution under monotonicity is difficult to determine, such as Loevinger's $H$ coefficient (Sijtsma & Molenaar, 2002, but see also Van der Ark, Croon & Sijtsma, 2008, who provide a test for $H_j$ under the marginal modeling approach). Bayesian alternatives require the specification of prior probabilities for the item probabilities and result in a test statistic for which the Type I error rate is difficult to establish (Karabatsos & Sheu, 2004). The current approach overcomes these difficulties by providing a likelihood ratio test for manifest monotonicity for individual items. This test has the added benefit of being sensitive to both global and local violations of latent monotonicity. By performing this test for every item and correcting for multiple testing, monotonicity can be evaluated for the test as a whole.

The next section presents the rationale behind the proposed approach. In the subsequent section, a likelihood ratio test is presented that is based on this approach, which can be used to determine whether manifest monotonicity should be rejected. The performance of this test is evaluated using a simulation study, in which the Type I error rate is investigated, as well as the power to detect violations of latent monotonicity. Finally, an empirical data example is used to illustrate the application of the testing procedure.

## 2.2 Order-Constrained Statistical Inference

A logical approach to testing latent monotonicity is to determine whether the data provide sufficient evidence to reject manifest monotonicity and, by implication, latent monotonicity. This focus on falsification is in line with the way in which model assumptions are traditionally evaluated. Thus, if the data do not suggest that manifest monotonicity is violated, the assumption of latent monotonicity can be retained, and one can consider applying IRT models that require this assumption to hold. For this purpose, a test to determine whether manifest monotonicity is violated is needed.

In the order-constrained statistical inference framework (see, e.g., Silvapulle & Sen, 2005), the current testing problem can be formulated as a Type B problem (Silvapulle & Sen, 2005), dealing with inequalities – manifest monotonicity – in the null hypothesis. Let $\pi_y = P(X = 1|Y = y)$, where the item index $i$ is dropped for convenience. Using this formulation, the null and alternative hypotheses correspond to:

$$\begin{aligned} H_0: &\quad \pi_0 \leq \pi_1 \leq \ldots \leq \pi_h, \quad \text{against} \\ H_1: &\quad H_0 \text{ does not hold.} \end{aligned}$$

If the data provide sufficient evidence that $H_0$ should be rejected in favor of $H_1$, manifest monotonicity is rejected.

With these formulations in mind, the procedures described by Silvapulle and Sen (2005) can be used to test $H_0$ against $H_1$. The application of these procedures in the context of testing manifest monotonicity is the topic of the following sections.

## 2.3 Likelihood Ratio Test

To determine whether the data indicate that manifest monotonicity is violated, a likelihood ratio test can be used to determine whether $H_0$ should be rejected in favor of $H_1$. To apply this test, three vectors containing maximum likelihood (ML) estimates of the conditional item probabilities in $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_h)$ are needed. These estimates are the maximum likelihood

(ML) estimates under three conditions: $\hat{\boldsymbol{\pi}}_e$ denotes the set of ML estimates when the conditional probabilities are required to be equal for all the levels of the manifest score, $\hat{\boldsymbol{\pi}}_m$ denotes the set of ML estimates when the probabilities are required to be in accordance with manifest monotonicity, and $\hat{\boldsymbol{\pi}}_u$ consists of the set of unconstrained ML estimates of the conditional probabilities.

These sets of estimates are easy to obtain, since the observations at the different levels of $Y$ can be assumed to originate from independent populations. The unconstrained ML estimates in $\hat{\boldsymbol{\pi}}_u$ are obtained through $\hat{\pi}_{uy} = \frac{s_y}{n_y}$, where $s_y$ denotes the number of positive responses (scores of 1) at $y$, and $n_y$ denotes the number of subjects with manifest score $y$. The estimates in $\hat{\boldsymbol{\pi}}_e$ can be obtained through

$$\hat{\pi}_{e0} = \cdots = \hat{\pi}_{eh} = \frac{\sum_{y=0}^{h} s_y}{\sum_{y=0}^{h} n_y}.$$

Thus, for every $y$, item probability $\hat{\pi}_{ey}$ is set equal to the overall proportion of positive responses.

To obtain $\hat{\boldsymbol{\pi}}_m$, a slightly adapted version of the *Pool Adjacent Violators Algorithm* (PAVA; see Silvapulle & Sen, 2005, p. 47) can be employed. This procedure always yields the ML estimates of $\boldsymbol{\pi}$ under manifest monotonicity. To start, the PAVA assigns the $h + 1$ estimated unconstrained conditional probabilities to $h + 1$ 'blocks', as displayed in Table 2.1. Moving from the first pair to the last pair, the algorithm checks whether for each pair of adjacent blocks the values in these blocks satisfy the constraints specified in Equation 2.3. If for each pair of adjacent blocks the constraints are satisfied, the estimates in the blocks are selected as $\hat{\boldsymbol{\pi}}_m$. If the constraints are not satisfied, then the estimates in at least two adjacent blocks show a decreasing order, which violates manifest monotonicity. The first two adjacent blocks that show a violation are then combined into one block, and the two estimates are combined into a weighted average. That is, if $\hat{\pi}_y > \hat{\pi}_{y+1}$, the updated estimate of both $\pi_y$ and $\pi_{y+1}$ becomes $\frac{s_y + s_{y+1}}{n_y + n_{y+1}}$, and they remain combined in a single block, reducing the total number of blocks by one. After this merger, the algorithm checks whether for all the

Table 2.1: Example of the application of PAVA.

| Iteration | $\hat{\pi}_0$ | | $\hat{\pi}_1$ | | $\hat{\pi}_2$ | | $\hat{\pi}_3$ | | $\hat{\pi}_4$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | $\{\frac{s_0}{n_0}\}$ | $>$ | $\{\frac{s_1}{n_1}\}$ | $>$ | $\{\frac{s_2}{n_2}\}$ | $<$ | $\{\frac{s_3}{n_3}\}$ | $>$ | $\{\frac{s_4}{n_4}\}$ |
| 1 | $\{\frac{s_0+s_1}{n_0+n_1}\}$ | | | $>$ | $\{\frac{s_2}{n_2}\}$ | $<$ | $\{\frac{s_3}{n_3}\}$ | $>$ | $\{\frac{s_4}{n_4}\}$ |
| 2 | $\{\frac{s_0+s_1+s_2}{n_0+n_1+n_2}\}$ | | | | | $<$ | $\{\frac{s_3}{n_3}\}$ | $>$ | $\{\frac{s_4}{n_4}\}$ |
| 3 | $\{\frac{s_0+s_1+s_2}{n_0+n_1+n_2}\}$ | | | | | $<$ | $\{\frac{s_3+s_4}{n_3+n_4}\}$ | | |
| $\hat{\boldsymbol{\pi}}_m$ | $\frac{s_0+s_1+s_2}{n_0+n_1+n_2}$ | $=$ | $\frac{s_0+s_1+s_2}{n_0+n_1+n_2}$ | $=$ | $\frac{s_0+s_1+s_2}{n_0+n_1+n_2}$ | $<$ | $\frac{s_3+s_4}{n_3+n_4}$ | $=$ | $\frac{s_3+s_4}{n_3+n_4}$ |

estimates the constraints in Equation 2.3 are now satisfied. If they are, the iterative process ends, and if they are not, again the first two adjacent blocks that show a decrease are merged and their estimates are averaged based on the number of observations in each block.

This merger may also take place between already combined blocks, for example, between the block of $y-1$ and the previously merged block of $y$ and $y+1$. The iterative process continues until the estimates in all the blocks satisfy the order constraints (i.e., the values are nondecreasing), after which the final values are selected as $\hat{\boldsymbol{\pi}}_m$. This way, estimates are obtained that are in accordance with manifest monotonicity, and these estimates are the ML estimates under Equation 2.3 (Silvapulle & Sen, 2005, pp. 45-46). An example of the application of the PAVA is displayed in Table 2.1, where a bold "$>$"-sign indicates which two blocks are merged during that iteration.

The values in $\hat{\boldsymbol{\pi}}_m$ are the ML estimates under $H_0$. For testing purposes, the values in $\hat{\boldsymbol{\pi}}_u$ can be used as the estimates under $H_1$. While using the unconstrained estimates $\hat{\boldsymbol{\pi}}_u$ does not exclude the possibility of having estimates under $H_1$ that are in accordance with manifest monotonicity, restricting the estimates in such a way that manifest monotonicity has to be violated has no added value for a test that aims to detect *violations* of manifest monotonicity. That is, if $\hat{\boldsymbol{\pi}}_u$ and $\hat{\boldsymbol{\pi}}_m$ are the same, there is no evidence in the data that manifest monotonicity is violated and it will not

be rejected, regardless of whether we do or do not restrict the estimates under $H_1$. By using the unconstrained ML estimates for $H_1$, the likelihood of those estimates is guaranteed to be at least as high as the likelihood of $\hat{\boldsymbol{\pi}}_m$, and this result ensures that a likelihood ratio test statistic based on those two sets of estimates is always nonnegative.

A likelihood ratio test can be constructed using the ML estimates in $\hat{\boldsymbol{\pi}}_m$ and $\hat{\boldsymbol{\pi}}_u$. Let $S_y$ be the random variable indicating the number of positive responses at $y$. Since each random variable $S_y$ has a binomial distribution with parameters $P(X = 1|Y = y)$ and $n_y$, and $S_0, \ldots, S_h$ are mutually independent, the log-likelihood is

$$\ell(\boldsymbol{\pi}|\boldsymbol{s}, \boldsymbol{n}) = \sum_{y=0}^{h} \ell(\pi_y|s_y, n_y),$$

where $\boldsymbol{s} = (s_0, \ldots, s_h)$, $\boldsymbol{n} = (n_0, \ldots, n_h)$, and

$$\ell(\pi_y|s_y, n_y) = s_y \log \pi_y + (n_y - s_y) \log(1 - \pi_y) + \log \left[ n_y! \{ s_y! (n_y - s_y)! \}^{-1} \right].$$

This way, the log-likelihood of both $\hat{\boldsymbol{\pi}}_m$ and $\hat{\boldsymbol{\pi}}_u$ can be obtained. The two log-likelihoods can be used to construct a likelihood ratio statistic,

$$T = -2 \left[ \ell(\hat{\boldsymbol{\pi}}_m|\boldsymbol{s}, \boldsymbol{n}) - \ell(\hat{\boldsymbol{\pi}}_u|\boldsymbol{s}, \boldsymbol{n}) \right].$$

This likelihood ratio statistic can be shown to be asymptotically $\bar{\chi}^2$-distributed (Silvapulle & Sen, 2005). To determine whether the likelihood ratio statistic obtained this way warrants the rejection of manifest monotonicity, the appropriate $\bar{\chi}^2$ null distribution needs to be obtained, which can be approximated through simulation.

Unlike traditional null hypotheses, $H_0$ does not specify point values for its parameters, but rather corresponds to a part of the parameter space $\boldsymbol{\pi} \in (0, 1)^h$. Thus, there is a wide range of values for $\hat{\boldsymbol{\pi}}_m$ that are allowed by $H_0$, and the distribution of $T$ under $H_0$ – the $T_0$-distribution, for short – depends on which values allowed by Equation 2.3 are selected. Thus, instead of a single distribution, a family of $T_0$-distributions exists – all asymptotically $\bar{\chi}^2$-distributed –, each of which results in a different value of $P(T_0 \geq T)$.

That is, the probability of observing a value of $T$ as extreme as the one observed in the data under the assumption that manifest monotonicity holds depends on which set of admissible values one uses as the reference point for the null hypothesis.

To avoid the possibility of inflating the Type I error rate, Silvapulle and Sen (2005, p. 91) suggest using the *least favorable null distribution*, which is the null distribution for which $P(T_0 \geq T)$ is maximized. Using this least favorable null distribution ensures that manifest monotonicity is not rejected if there is at least one realization of $\boldsymbol{\pi}_m$ that is consistent with the data. Thus, when the nominal significance level is set to $\alpha$, the Type I error rate never exceeds $(100 \times \alpha)\%$.

Whenever values of $\hat{\boldsymbol{\pi}}_m$ are used that are not all equal (but still ordered in accordance with manifest monotonicity), the probability of observing a value of $T_0$ that exceeds a certain value can always be increased by using $\hat{\boldsymbol{\pi}}_e$ instead of $\hat{\boldsymbol{\pi}}_m$. This follows from the fact that the values that are admissible for $\hat{\boldsymbol{\pi}}_e$ are also admissible values for $\hat{\boldsymbol{\pi}}_m$, but not vice versa. Hence, the likelihood of $\hat{\boldsymbol{\pi}}_m$ can never be lower than the likelihood of $\hat{\boldsymbol{\pi}}_e$. Thus, the least favorable null distribution of $T$ under $H_0$ is obtained when $\hat{\boldsymbol{\pi}}_m$ is replaced by $\hat{\boldsymbol{\pi}}_e$, and the estimates $\hat{\boldsymbol{\pi}}_e$ are located at the boundary of the subspace that corresponds to $H_0$ (Silvapulle & Sen, 2005, p. 91). By using $\hat{\boldsymbol{\pi}}_e$, the least favorable null distribution of $T$ is obtained, ensuring that the actual significance level can never exceed the nominal significance level.

By repeatedly simulating data using $\hat{\boldsymbol{\pi}}_e$, the least favorable null distribution of $T$ can be approximated to any degree of precision. Using this distribution, the probability of observing a value of $T$ under this null situation that is at least as extreme as the one observed in the data – $P(T_0 \geq T)$ – can be approximated. If this $p$-value is not lower than the specified $\alpha$, then there is at least one set of values admissible by Equation 2.3 with which the data are consistent, and manifest monotonicity cannot be rejected. If this value is lower than $\alpha$, then there is no set of values for $\boldsymbol{\pi}_m$ which with the data are consistent, and both manifest and latent monotonicity are rejected for the item under investigation.

Problems may arise when the conditional item probabilities have values close to 0 or 1, or when the number of observations is small at certain lev-

els of the manifest score. In those cases, some of the simulated data sets may have observed proportions equal to 0 or 1. Since the corresponding estimated conditional probabilities would then also equal 0 or 1, the likelihood ratio statistics cannot be calculated, because this requires evaluating both $\log \hat{\pi}_y$ and $\log(1 - \hat{\pi}_y)$ for every $y$. This problem could also hold for the calculation of $T$ for empirical data if some of the observed conditional proportions are 0 or 1. If this problem arises at level $y$, it may be solved by merging that level with an adjacent level of the manifest score, $y - 1$ or $y + 1$. Since there is no reason to prefer a merger with either the previous or the next level, the testing procedure can be set to merge a level $y$ with the adjacent level that has the smaller number of observations. This way, relatively little information with regard to manifest monotonicity will be lost.

If the observed proportions for the two levels that are merged differ from each other, merging these two levels solves the calculation problem. If the two observed proportions are both 0 or both 1, another adjacent level of the manifest score should be merged with the previous two levels. This procedure continues until none of the observed proportions is equal to 0 or 1, or until only one level of the manifest score remains. In the latter case, $T$ should be set to 0, since there is no evidence against monotonicity when the manifest score has only one level, but this is obviously a trivial case in which the data are too sparse to be informative about latent monotonicity.

While some information concerning manifest monotonicity may be lost when adjacent levels have to be merged due to problems with sparsity, this loss can generally be assumed to be minimal since mergers are mainly necessary for manifest scores with few observations, which would not contribute much information. It is however a clear indication that having sufficiently many observations at the different levels of the manifest score is to be preferred over dealing with small samples. Of course, one could also consider applying some other form of sparsity correction or a smoothing algorithm to ensure that the estimated conditional probabilities are never equal to 0 or 1. Since the information loss due to the correction we suggested is expected to be minimal, we did not pursue these alternative approaches.

## 2.4 Simulation Study

### 2.4.1 Method

A simulation study was performed to evaluate the Type I error rate of the proposed test and its power to detect violations of manifest monotonicity. To evaluate the Type I error rate, the rejection rate has to be evaluated under $H_0$. Similarly, the power of the test to detect violations of manifest monotonicity can be evaluated by investigating the rejection rate of the test under $H_1$. Since both hypotheses specify a part of the parameter space rather than specific point values for the parameters, choices have to be made as to which values are used for both conditions. Rather than specifying arbitrary values for the conditional probabilities at the manifest level, item response functions (IRFs) that correspond to either the $H_0$ or the $H_1$ situation were specified for this simulation study. This way, the values resulting at the manifest level are less arbitrary and easier to interpret, since they result from functions specified at the latent level.

For evaluating the Type I error rate, a natural choice is to focus on the least favorable null situation, since this is the situation for which the level of significance is assumed to be fixed. Any other choice should result in a rejection rate that is *lower* than the specified level of significance. Thus, the IRF of the item that is investigated under the null situation should be constant with respect to the latent trait; we chose a constant IRF value of .5. For this item the probability of obtaining a positive score is independent of the latent variable. This ensures that the probability of obtaining a positive score on the item conditional on the manifest score is also constant, which corresponds to the least favorable null situation.

For evaluating the power of the test to detect violations of manifest monotonicity, a nonmonotone IRF has to be specified. Unlike the null situation, there is not one obvious choice for the IRF. Any selected IRF will be arbitrary to some extent, but we have opted for selecting an IRF that shows a moderate violation of monotonicity around the center of the distribution of the latent variable. For this moderate violation we chose an IRF that showed a decline in probability of about .4 over a range of about 2 stan-

dard deviations on the latent variable distribution. Such an IRF might, for example, be thought to correspond to an item from developmental psychology, measuring the performance on some task that is affected when children at one point acquire a new skill that is at first difficult to apply (resulting in a decline of the IRF) but which ends up being useful after it is fully mastered (e.g., the acquisition of a new grammatical rule, which tends to be overused when it is first mastered).

A nonmonotone IRF was obtained using a simple polynomial extension of the two-parameter logistic model,

$$P(X_i = 1|\theta) = \frac{exp\left(a_{1i}(\theta - b_{1i}) + a_{2i}(\theta - b_{2i})^2 + a_{3i}(\theta - b_{3i})^3\right)}{1 + exp\left(a_{1i}(\theta - b_{1i}) + a_{2i}(\theta - b_{2i})^2 + a_{3i}(\theta - b_{3i})^3\right)},$$

where $b_{1i}$, $b_{2i}$ and $b_{3i}$ function as the item's location parameters for the different-order polynomials, and $a_{1i}$, $a_{2i}$ and $a_{3i}$ determine their slope and hence the extent to which they determine the shape of the IRF. By setting $a_{2i}$ and $a_{3i}$ to 0, the two-parameter logistic model is obtained. Here, the purpose of this model was to obtain a useful nonmonotone IRF, and hence issues of estimation and identification are irrelevant for this study. This model can be used to produce a variety of IRFs with a local violation of latent monotonicity. To obtain an IRF that shows the moderate violation that was specified above, we set $a_{1i}$, $a_{2i}$ and $a_{3i}$ to 1, 1.2 and .25, and $b_{1i}$, $b_{2i}$ and $b_{3i}$ were set to 2.5, 1.6 and 1.5, respectively.

In addition to the item with a local violation of latent monotonicity, two items were considered that showed a global violation of latent monotonicity. The IRFs of these items are monotonely decreasing, and hence violate latent monotonicity across the full range of the latent variable. The first item was constructed using the two-parameter logistic (2PL) model, where the difficulty parameter was set to 0 and the discrimination parameter was set to -1. This item can be thought of as an item measuring the latent variable that was incorrectly coded. The second item was constructed using a four parameter logistic (4PL) model, again with a difficulty parameter of 0, and discrimination parameter of -1, but with a lower asymptote of .25 and an upper asymptote of .75. Such an item is still somewhat related to the construct that is intended to be measured, but its relation being

in a different direction than the researcher would initially have expected. This item could thus be considered to be a 'surprising' item, one that has not simply received the wrong coding, but actually does not behave in accordance with expectations. The simulation study focused on the rejection rates that were observed when manifest monotonicity was tested for each of these four items individually. Figure 2.1 displays the shape of the IRFs of these four items.

In order to obtain a manifest score over which monotonicity can be evaluated, other items were needed. A set of five items was constructed using the two-parameter logistic model. The difficulty of the items ranged from -1 to 1, and their discrimination parameter ranged from .5 to 1.5. Figure 2.2 shows the shapes of these IRFs. To evaluate the effect of test length on the power of the statistical test, sets of 10 and 20 items were also considered, and these sets consisted of two or four times the original set of items with a monotone IRF. In addition to the number of items, the sample size was also varied, by using values of 100, 200, 300, 400 and 500.

For each of the specified numbers of monotone items (denoted by $k$) and each sample size (denoted by $n$), 2000 data sets were generated. This was done by drawing $n$ values from a standard normal distribution for each replication, and using these draws as values on the latent variable. These values were used to calculate the probability of observing a positive score on each of the monotone items, as well as the probability for the item under investigation (one of the four irregular items described earlier). Using these probabilities, data were generated by drawing values from a Bernoulli distribution. For each data set, the $T$-statistic was calculated. After obtaining this statistic, the corresponding null distribution was approximated for each simulated data set by generating 2000 data sets using the $\hat{\boldsymbol{\pi}}_e$ and $\boldsymbol{n}$ for that data set, resulting in an approximation of the $p$-value for that replication. If the obtained $p$-value was smaller than .05, then this constituted a rejection of manifest monotonicity for that data set. The percentage of rejections was determined for each combination of $k$ and $n$. This was done separately for each of the four irregular items, and in that way the Type I error rate and the power of the test were evaluated.

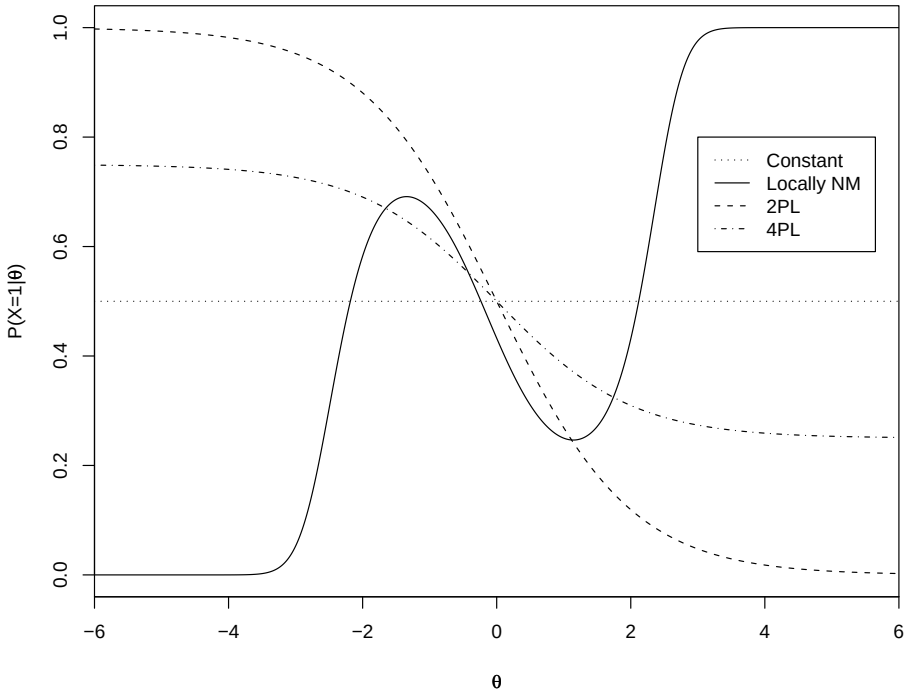


Figure 2.1: The item characteristic curves of the four items that were analyzed. The constant function is denoted by 'Constant', the locally non-monotone function is denoted by 'Locally NM', and the two-parameter and four-parameter logistic functions are denoted by '2PL' and '4PL', respectively.

### 2.4.2 Results

The results of this simulation study are displayed in Table 2.2. As can be observed, when manifest monotonicity was evaluated for the item with the

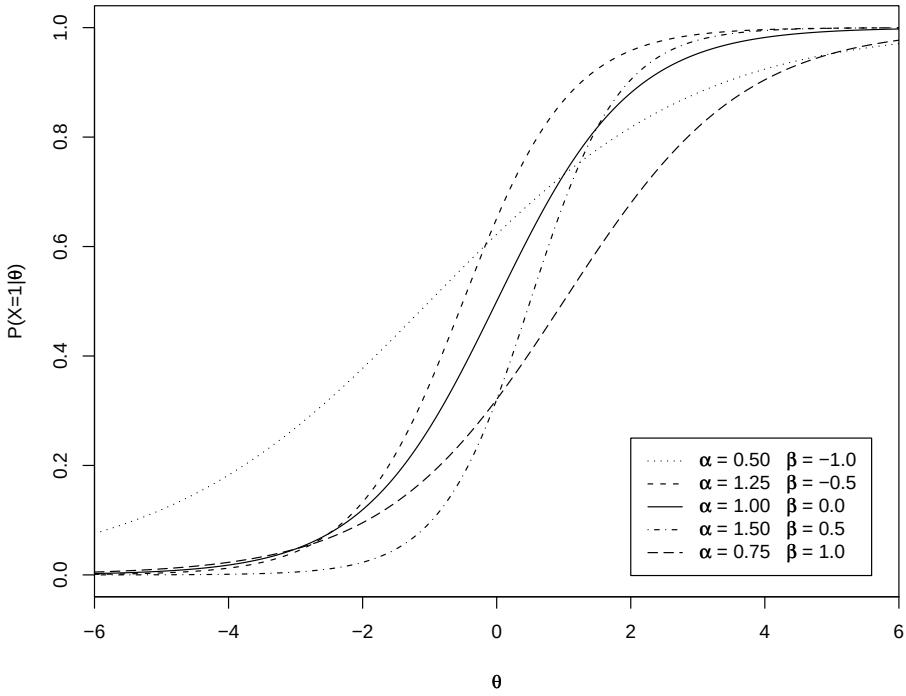


Figure 2.2: The item characteristic curves of the five monotone items, based on the two-parameter logistic model. The discrimination and difficulty parameters are denoted by $\alpha$ and $\beta$, respectively.

constant probability, the rejection rate was close to .05 for all conditions. The rejection rate does not appear to depend on either sample size or the number of items included in the manifest score. Thus, there is no indication that the Type I error rate under the least favorable null situation deviates from the specified level of significance.

Table 2.2 also displays the proportion of rejections for the different

Table 2.2: Rejection rates of manifest monotonicity for the items with constant IRF and nonmonotone IRFs (2,000 replications).

| | Constant IRF | | | Locally nonmonotone IRF | | |
|---|---|---|---|---|---|---|
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .055 | .049 | .059 | .334 | .404 | .457 |
| 200 | .051 | .051 | .058 | .553 | .645 | .652 |
| 300 | .048 | .050 | .050 | .741 | .800 | .842 |
| 400 | .050 | .055 | .045 | .857 | .904 | .930 |
| 500 | .049 | .043 | .053 | .931 | .956 | .980 |

| | Decreasing 2PL IRF | | | Decreasing 4PL IRF | | |
|---|---|---|---|---|---|---|
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .754 | .804 | .845 | .272 | .278 | .328 |
| 200 | .963 | .982 | .989 | .440 | .466 | .446 |
| 300 | .997 | 1.000 | 1.000 | .621 | .638 | .605 |
| 400 | 1.000 | 1.000 | 1.000 | .758 | .766 | .767 |
| 500 | 1.000 | 1.000 | 1.000 | .836 | .846 | .826 |

conditions for the item with a local violation of latent monotonicity. The rejection rate depends heavily on the sample size, with a larger sample size corresponding to higher power. Additionally, the power of the test is positively related to the number of items on the test. For tests with either 10 or 20 items, a power level of about .80 is reached when a sample size of 300 was used, while a test with 5 items would require a somewhat larger sample size.

For the two items with a global violation of latent monotonicity, the rejection rate of manifest monotonicity also depends heavily on sample size. Additionally, it can be observed that for every condition the rejection rate for the 2PL-model item is much higher than that of the 4PL-model

item. For the 2PL item, power levels exceeding .80 were observed for all conditions, except when $k = 5$ and $n = 100$. For the 4PL item, these power levels are only observed when $n = 500$, and this did not depend on the number of items on the test. Even when the 4PL-model item was evaluated using a sample size of 200, the rejection rate still exceeded .90 for all test lengths. These results indicate that even for small sample sizes the procedure is quite capable of rejecting manifest monotonicity when a monotonely decreasing IRF is considered, and hence that it easily detects global violations of latent monotonicity.

Two things may be noted. First, since the Type I error rate is fixed using the least favorable null situation, the Type I error rate is smaller than $\alpha$ for more favorable null situations, that is, when an item is considered with a monotonely increasing IRF. For example, when the same simulation procedure was applied to a 2PL-model item with a discrimination parameter of 1 and a difficulty parameter equal to the population mean of the latent variable, the rejection rate was close to zero for all conditions (results not displayed). By fixing the Type I error rate for the least favorable null situation, the test automatically becomes conservative for more favorable null situations.

It was already noted that other null situations might result in a rejection rate that is lower than $\alpha$, but for small sample sizes some situations corresponding to $H_1$ can likewise result in rejection rates that are lower than $\alpha$. If the conditional probabilities show a strong positive trend, this trend can mask small local deviations from monotonicity. In such a case, most conditional probabilities show a monotone ordering, resulting in a value for $T$ that could on average be *lower* than one would expect under the least favorable null situation. However, as the sample size increases the deviations that are expected under the null situation become smaller and smaller, up to the point where even a small deviation from monotonicity in the observed proportions is highly unlikely under the null situation, leading to a rejection of manifest monotonicity. Thus, the power to detect violations of monotonicity should tend to 1 as the sample size increases, which is supported by the findings displayed in Table 2.2. While power levels lower than $\alpha$ were not observed in this simulation study, this may cause a

problem when trying to detect very small violations of monotonicity.

To illustrate the extent to which a local violation of monotonicity at the latent level (as displayed in Figure 2.1) results in a violation of monotonicity at the manifest level, the probabilities at the manifest level for $k = 5$ and $k = 10$ for the locally nonmonotone item are shown in Figure 2.3 and Figure 2.4, respectively. It can be observed that a moderate violation of latent monotonicity might result in a violation of monotonicity at the manifest level that is only minor if the number of items included in the manifest score is small. When violations of latent monotonicity are considered that are even smaller, the violation at the manifest level may disappear entirely if one does not include enough items in the manifest score, rendering it impossible to detect the violation of monotonicity at the latent level. Thus, to have sufficient power to detect violations of latent monotonicity of smaller sizes than the ones considered here, it is crucially important to include a sufficiently large number of items in the manifest score.

## 2.5 Empirical Example

We applied the testing procedure to every item from a scale measuring nonaggressive antisocial behavior in male youths (Dekovic, 2003) to test for evidence against manifest monotonicity, and hence to determine whether the application of the monotone homogeneity model might be appropriate. In order to evaluate latent monotonicity for the scale as a whole, the lowest observed $p$-value should be contrasted with its critical value, after correcting the $\alpha$-level for multiple testing. The scale consisted of seven items, asking the respondent for the frequency with which they performed specific types of nonaggressive antisocial behavior during the previous year, such as disregarding orders from parents or shoplifting. Since most of these behaviors were rare, the item frequency distributions were highly skewed. Therefore, the items were dichotomized by assigning a value of 1 if the behavior did occur during that year and a value of 0 if it did not occur.

Monotonicity was evaluated across the restscore, and the level of significance was set at .05. Selecting the restscore is an attractive choice, since

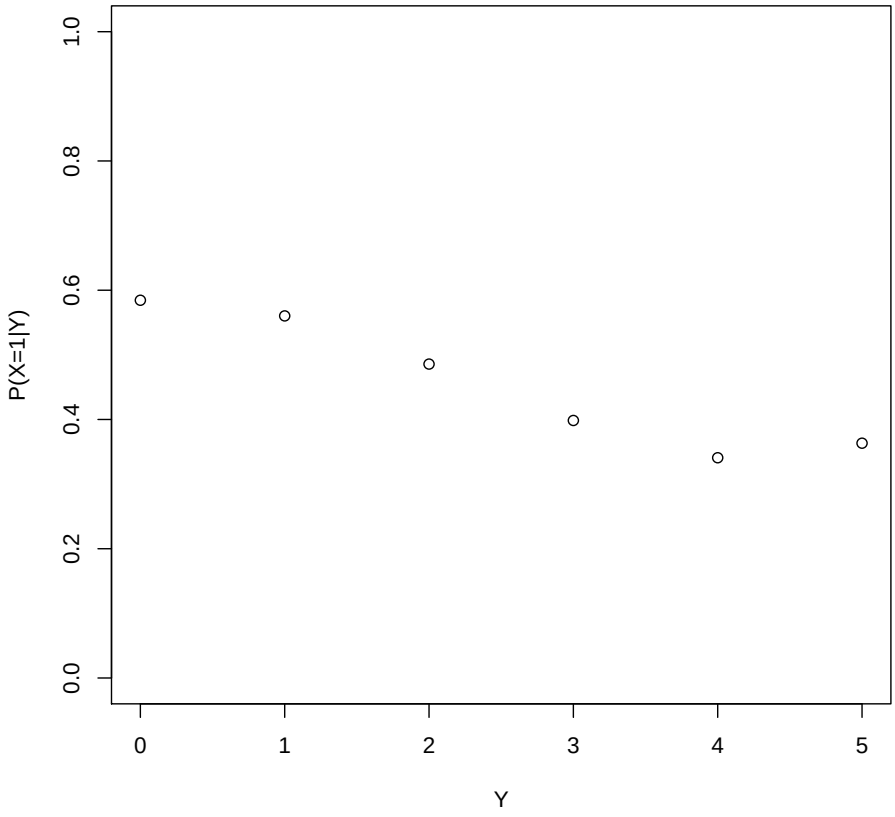


Figure 2.3: Conditional probabilities for the locally nonmonotone item with 5 items included in the manifest score.

it is a more reliable ordinal estimator of the latent variable than manifest scores based on fewer items. Table 2.3 displays the observed proportions of positive scores for each of these items and for each of the restscore groups.

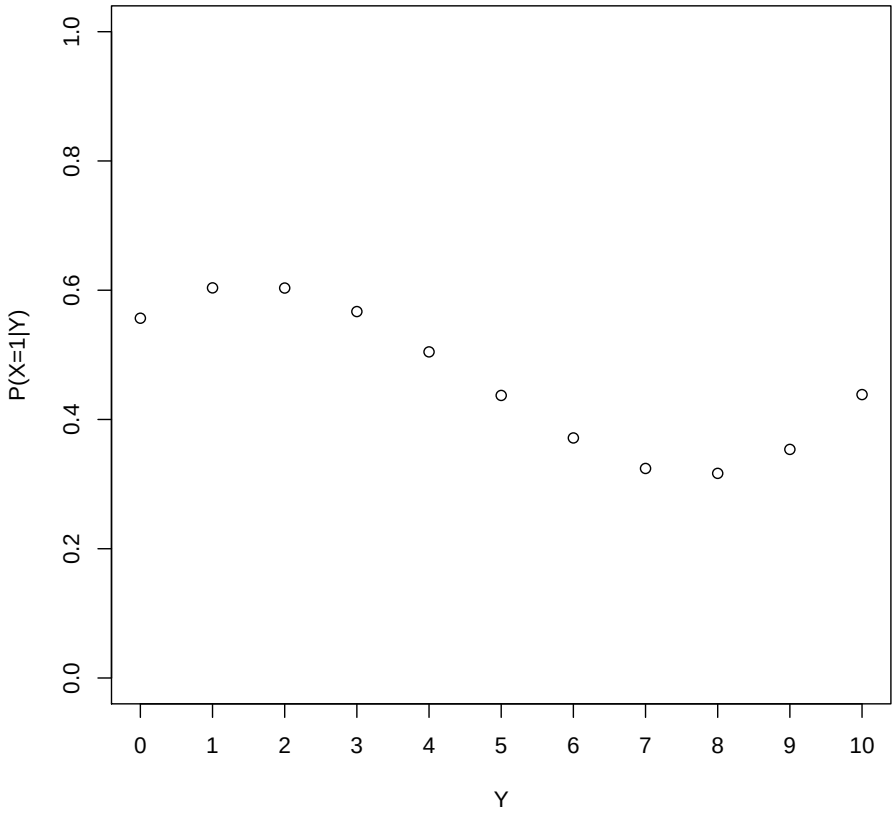


Figure 2.4: Conditional probabilities for the locally nonmonotone item with 10 items included in the manifest score.

Discordant orderings of the proportions are indicated by a "$>$"-sign. Most of the observed proportions were ordered in accordance with manifest monotonicity. Except for item 7, all the items showed at most one discordant

Table 2.3: Observed conditional proportions for the nonaggressive antisocial behavior data.

| Item | $\frac{s_0}{n_0}$ | | $\frac{s_1}{n_1}$ | | $\frac{s_2}{n_2}$ | | $\frac{s_3}{n_3}$ | | $\frac{s_4}{n_4}$ | | $\frac{s_5}{n_5}$ | | $\frac{s_6}{n_6}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1: Disregarding parents | .49 | $<$ | .66 | $<$ | .75 | $<$ | .88 | $>$ | .84 | $<$ | .97 | $<$ | 1.00 |
| 2: Missing a curfew | .19 | $<$ | .29 | $<$ | .43 | $<$ | .64 | $<$ | .74 | $<$ | .90 | $<$ | 1.00 |
| 3: Skipping school | .02 | $<$ | .08 | $<$ | .13 | $<$ | .17 | $<$ | .25 | $<$ | .55 | $>$ | .33 |
| 4: Cheating on a test | .18 | $<$ | .34 | $<$ | .41 | $<$ | .55 | $<$ | .69 | $<$ | .96 | $>$ | .75 |
| 5: Fare dodging | .08 | $<$ | .17 | $<$ | .19 | $<$ | .35 | $<$ | .40 | $<$ | .69 | $<$ | .75 |
| 6: Shoplifting | .04 | $<$ | .14 | $<$ | .18 | $<$ | .29 | $<$ | .39 | $<$ | .78 | $>$ | .60 |
| 7: Stealing | .04 | $>$ | .04 | $<$ | .11 | $<$ | .14 | $<$ | .18 | $<$ | .53 | $>$ | .40 |

Table 2.4: Estimated conditional probabilities for item 7.

| | $y$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimates | 0 | | 1 | | 2 | | 3 | | 4 | | 5 | | 6 |
| $\hat{\boldsymbol{\pi}}_u$ | .04 | $>$ | .04 | $<$ | .11 | $<$ | .14 | $<$ | .18 | $<$ | .53 | $>$ | .40 |
| $\hat{\boldsymbol{\pi}}_m$ | .04 | $=$ | .04 | $<$ | .11 | $<$ | .14 | $<$ | .18 | $<$ | .49 | $=$ | .49 |
| $\hat{\boldsymbol{\pi}}_e$ | .14 | $=$ | .14 | $=$ | .14 | $=$ | .14 | $=$ | .14 | $=$ | .14 | $=$ | .14 |

ordering of the proportions. Most of these discordant orderings concerned the two highest restscore groups, which contained relatively few observations. For Items 1 and 2, a proportion of 1 was found for the highest restscore group. Thus, to deal with sparsity, the last two restscore groups were merged for the analysis of these two items.

The testing procedure was applied separately to each of the seven items. To illustrate the application of the procedure, the analysis of Item 7 – the item with the highest number of discordant orderings – is discussed in more detail. Table 2.4 displays the estimated probabilities for this item, and Table 2.5 shows how the estimates of $\boldsymbol{\pi}_m$ are obtained through the application of the PAVA.

Table 2.5: Application of PAVA to item 7.

| | $y$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Iteration | 0 | | 1 | | 2 | | 3 | | 4 | | 5 | | 6 |
| 0 | $\{\frac{3}{72}\}$ | $>$ | $\{\frac{4}{111}\}$ | $<$ | $\{\frac{13}{123}\}$ | $<$ | $\{\frac{11}{81}\}$ | $<$ | $\{\frac{10}{56}\}$ | $<$ | $\{\frac{20}{38}\}$ | $>$ | $\{\frac{6}{15}\}$ |
| 1 | | | $\{\frac{7}{183}\}$ | $<$ | $\{\frac{13}{123}\}$ | $<$ | $\{\frac{11}{81}\}$ | $<$ | $\{\frac{10}{56}\}$ | $<$ | $\{\frac{20}{38}\}$ | $>$ | $\{\frac{6}{15}\}$ |
| 2 | | | $\{\frac{7}{183}\}$ | $<$ | $\{\frac{13}{123}\}$ | $<$ | $\{\frac{11}{81}\}$ | $<$ | $\{\frac{10}{56}\}$ | $<$ | | $\{\frac{26}{53}\}$ | |
| $\hat{\boldsymbol{\pi}}_m$ | .04 | $=$ | .04 | $<$ | .11 | $<$ | .14 | $<$ | .18 | $<$ | .49 | $=$ | .49 |

The log-likelihoods of $\hat{\boldsymbol{\pi}}_m$ and $\hat{\boldsymbol{\pi}}_u$ were $-13.27$ and $-12.90$, respectively, resulting in $T = .73$. The null distribution for this statistic was approximated by repeatedly generating data using $\hat{\boldsymbol{\pi}}_e$ and $\boldsymbol{n}$. The $p$-value was approximated using 50,000 replicated sets of data. A value of .929 was obtained, indicating that the data do not provide sufficient evidence to warrant a rejection of manifest monotonicity. Figure 2.5 shows the obtained null distribution of $T$, where the observed value of $T$ is indicated by a dashed vertical line.

The results for the other items were similar; the estimated $p$-values for each of the items indicated that there is no reason to reject monotonicity, since the lowest $p$-value – .562, obtained for Item 4 – does not lead to a rejection regardless of which type of correction for multiple testing one selects. Thus, there is no reason to reject latent monotonicity for the items in this scale, and this corroborates the assumption that subjects can appropriately be ordered using their total scores.

## 2.6 Discussion

Latent monotonicity is an assumption that is shared by most dichotomous IRT models. The statistical approach discussed in this paper provides a way of evaluating this assumption, by determining whether the data indicate that manifest monotonicity is violated or whether it can be maintained.

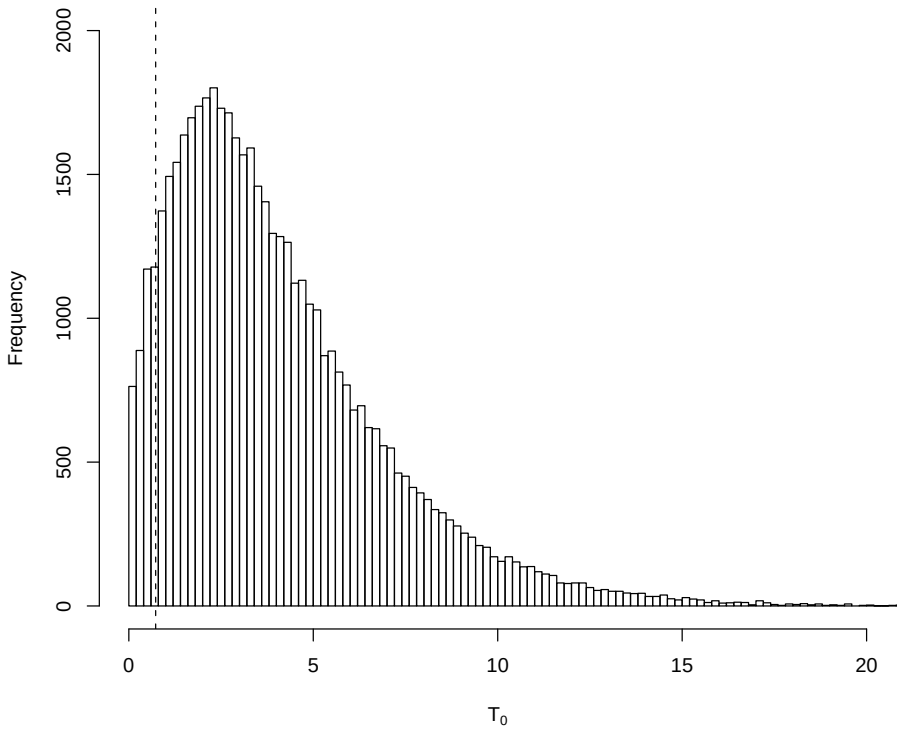


Figure 2.5: Distribution of $T_0$ for Item 7, obtained through simulation. The dashed line indicates the value of $T$ in the sample.

Since latent monotonicity and local independence together imply manifest monotonicity over the rest score, a rejection of manifest monotonicity over the rest score can be seen as a rejection of the monotone homogeneity model and its special cases. Thus, evaluating manifest monotonicity can be seen as a first step that one can take before assessing more specific IRT models.

The current approach treats latent monotonicity as an assumption that

should be evaluated by translating it into a null hypothesis that can be tested. This should make the procedure well-suited for practical applications, where latent monotonicity is often assumed to hold. Another approach would be to search for evidence in favor of monotonicity, thus aiming at confirmation rather than falsification. However, such an approach could end up overlooking violations of monotonicity if there is an overall positive trend, similar to the locally nonmonotone item that was used in the simulation study. The proposed procedure is able to detect both local and global violations of monotonicity, thus avoiding this problem. An additional benefit of the procedure is that the Type I error rate is controlled under the least favorable null situation, ensuring that the probability of incorrectly rejecting manifest monotonicity does not exceed the specified level of significance. Thus, the procedure provides a test with a known Type I error rate for determining whether the monotone homogeneity model fits an item.

In order to evaluate whether applying the monotone homogeneity model to a set of items may be appropriate, the proposed test can be applied to each item separately. If manifest monotonicity is retained for every item after correcting for multiple testing, this supports the claim that the model holds for the set of items. If there is at least one significant result after correcting for multiple testing, the model should be rejected for the complete item set. It should be noted that since the approach focuses on monotonicity over a manifest score, whether monotonicity is rejected for a specific item may depend on which items are included in the manifest score. For this reason, we recommend removing the item with the lowest significant $p$-value and afterwards re-evaluating monotonicity for the remaining items, instead of blindly eliminating all items with a significant deviation from monotonicity from the item set. Regardless of which specific procedure for item selection and removal is followed, care should be taken to prevent an inflation of the Type I error rate. Sijtsma and Molenaar (2002) discuss the topic of item selection for the monotone homogeneity model more extensively.

The power values observed in our simulation study were high, indicating that the test can be powerful in detecting violations of latent monotonicity. This power depends to a large extent on the sample size and the severity

of the violation, and to a smaller extent on the number of items included in the manifest score. When only a small number of items is used, each level of the manifest score covers a relatively broad range of values on the latent variable, which reduces the power to detect local violations of latent monotonicity. Thus, in order to have a powerful check as to whether the assumption of latent monotonicity can be maintained, a sufficiently large number of items should be used. With regard to sample size, the estimated power values exceeded .80 for all types of violations when $n = 500$, which suggests that the test can successfully be used to detect both these types of violations for reasonable sample sizes.

The procedure requires the selection of a specific point in the null parameter space as the null value to be used in obtaining a null distribution for the test statistic. If one wants to avoid specifying such a null value, Bayesian alternatives could be considered. For example, Karabatsos and Sheu (2004) proposed a procedure that uses a Gibbs sampler to determine whether it is likely that manifest monotonicity over the restscore holds. This procedure could also be applied to other manifest scores. It is worth investigating the possible benefits and drawbacks of such a procedure compared to the procedure discussed in this article. It would be interesting to compare the power of their approach and the current approach. However, there are substantial conceptual differences between these approaches that have to be taken into consideration if one aims to make such a comparison. For example, the current approach uses a traditional $p$-value and deliberately fixes the Type I error rate for the least favorable null situation, whereas the Bayesian approach uses a posterior predictive $p$-value and considers the entire parameter space.

Additionally, it is worth considering extensions of the proposed procedure, for example to polytomous items, to investigate whether monotonicity of the expected item score over some manifest score can be maintained. Another interesting extension is the adaptation of the procedures to the evaluation of invariant item ordering, one of the assumptions of the double monotonicity model (Mokken, 1971). Since invariant item ordering implies that the item probabilities conditional on a variety of manifest scores (Rosenbaum, 1987b; Tijmstra, Hessen, Van der Heijden & Sijtsma, 2011)

have the same weak order, the order-constrained statistical inference framework may provide a useful approach to evaluating this property.

# Chapter 3

# Evaluating Manifest Monotonicity Using Bayes Factors

The assumption of latent monotonicity in item response theory cannot be evaluated directly, but it implies manifest monotonicity across a variety of observed scores such as the restscore, facilitating the assessment of latent monotonicity in real data. Standard methods of evaluating manifest monotonicity typically produce a test statistic that is geared towards falsification, which can only provide indirect support in favor of manifest monotonicity. We propose an alternative Bayesian method that more directly captures the amount of support for manifest monotonicity available in the data. Through the use of Bayes factors, the support for manifest monotonicity can be quantified. By using informative alternative hypotheses in addition to using the complement of manifest monotonicity, this procedure is also able to determine the support in favor of manifest monotonicity over for example the possibility that monotonicity does not hold for high or low ability respondents, or the presence of local violations of monotonicity. This possibility of examining different alternative hypotheses makes the procedure highly flexible. The performance of the procedure is evaluated using a

simulation study, and the application of the procedure is illustrated using empirical data.

## 3.1 Introduction

In dichotomous item response theory (IRT), one of the main assumptions shared by most parametric and non-parametric response models is that of latent monotonicity. This assumption states that the probability of observing a positive response to an item increases monotonically as the latent variable increases, and it plays an important role in obtaining the monotone likelihood-ratio property of the total score (Hemker, Sijtsma, Molenaar & Junker, 1997). The monotone likelihood-ratio property ensures that the total score can be used to order respondents on the latent variable, and this ordinal level of measurement is crucial to most applications of IRT. The assumption of latent monotonicity also captures the idea that the items on a test measure the latent variable (Junker & Sijtsma, 2000). For these reasons, investigating whether the assumption of latent monotonicity holds is important and relevant for many applications of IRT.

Since the latent variable is not observable, the assumption of latent monotonicity can only be evaluated indirectly, by considering observed item responses. Under the assumption of local independence, latent monotonicity implies monotonicity over a variety of manifest scores, such as a single item score (Mokken, 1971), the unweighted restscore (Rosenbaum, 1984; Junker & Sijtsma, 2000), or any sumscore that does not include the item in question. By testing whether monotonicity holds at the manifest level, one can investigate (under the assumption of local independence) whether latent monotonicity is violated. Tijmstra, Hessen, Van der Heijden and Sijtsma (2013) showed how the property of manifest monotonicity can be evaluated for a variety of manifest scores using order-constrained statistical inference, resulting in a likelihood-ratio test that determines whether there is sufficient evidence to reject monotonicity for the manifest score that was chosen. A rejection of manifest monotonicity implies a violation of latent monotonicity, so a significant test statistic results in a rejection

of latent monotonicity. Other methods of investigating latent monotonicity have been proposed, making use of either a manifest score (see, e.g., Rosenbaum, 1984) or the set of observed response patterns (Scheiblechner, 2003). Other nonparametric approaches have been developed that try to estimate the item characteristic curve, for example through the use of kernel smoothing (Ramsay, 1991) and spline-fitting (Abrahamowicz & Ramsay, 1992), where local statistical tests and confidence bands can be used to assess manifest monotonicity.

However, one of the issues that arises in the aforementioned approaches to evaluating latent monotonicity is that the tests based on these approaches all have to make use of a null hypothesis that specifies a boundary case of manifest monotonicity. That is, the null hypothesis that is tested in each of these approaches is the 'least favorable null hypothesis' (Silvapulle & Sen, 2005) that still corresponds to manifest monotonicity, which is tested against the alternative hypothesis that manifest monotonicity does not hold. The specific form of this null hypothesis differs for each of these approaches, but they all make use of the boundary case where there is no association between the item scores and hence where the item probabilities neither increase nor decrease over the manifest score. The rationale behind using this hypothesis is that it considers the boundary of the part of the parameter space that corresponds to manifest monotonicity: if manifest monotonicity cannot be rejected for those parameter values, the data are consistent with at least one point in the parameter space that corresponds to manifest monotonicity. However, since in test construction items are usually designed to measure a common latent variable, this null hypothesis is highly implausible in most practical settings where one would want to evaluate monotonicity.

Although these approaches are theoretically sound, in making use of the least favorable null hypothesis they may end up losing power to detect violations of manifest monotonicity. That is, in controlling the Type I error rate and ensuring that it does not exceed the specified level of significance and that latent monotonicity is not rejected if there is at least one point in the parameter subspace with which the data are consistent, these approaches may be erring on the conservative side and end up inflat-

ing the Type II error rate – that is, fail to accumulate enough evidence to correctly reject latent monotonicity. Failing to detect violations of latent monotonicity could lead to using an IRT model whose estimates cannot be trusted. Arguably, this could be worse than incorrectly concluding that latent monotonicity does not hold and not applying an IRT model. Thus, it is crucially important that a test for latent monotonicity has sufficient power to detect violations.

Furthermore, the approaches discussed so far make use of the null hypothesis testing framework, with an orientation that is explicitly aimed at falsification. That is, the tests attempt to provide a 'critical test' for the model assumption, to see whether the assumption is able to 'survive' this test. However, failing to reject an assumption does not imply that it actually holds, since a Type II error could have been made. Since model assumptions have to hold for the model to be valid, one can argue that simply noting that the assumption has failed to be rejected does not suffice as justification for applying the model. Of course, performing a power analysis may help to some extent to indirectly assess the amount of support that the model assumption receives when it fails to be rejected. However, it could be argued that a more direct way of assessing support *in favor* of the model assumption is needed if a decision needs to be made about whether we are justified to apply the model, which the discussed frequentist approaches do not provide.

It is with these goals of increasing the power and directly assessing the support *in favor* of monotonicity in mind that developing a different way of testing latent monotonicity may be highly relevant. Instead of using methods based on traditional null hypothesis testing, an alternative would be to consider a Bayesian model comparison approach. There are many different Bayesian model comparison approaches available (see for example Gelman, Carlin, Stem & Rubin, 2004), but of special interest here is the approach that focuses on the Bayes factor (see Kass & Raftery, 1995, or Hoijtink, 2012). Under such an approach, different hypotheses may be compared without giving a special status to one of the hypotheses by labeling it as a 'null hypothesis', and rather than attempting to reject this null hypothesis, one investigates which hypothesis receives the most support from the data.

Also, rather than resulting in a dichotomous outcome – reject or retain the assumption of latent or manifest monotonicity based on whether a $p$-value is obtained that is considered to be significant –, such an approach would be able to quantify the degree of support each hypothesis receives from the data. This approach could give researchers more information about the plausibility of the different hypotheses and would enable them to make an informed decision about the credibility of the assumption of latent monotonicity and whether they are confident enough to accept that this assumption holds. Furthermore, such a Bayesian approach would allow for more than just contrasting the hypothesis of manifest monotonicity with the general hypothesis that manifest monotonicity does not hold (Tijmstra et al., 2013). Rather, a wide variety of hypotheses that could be deemed relevant in the context of monotonicity could be compared, allowing for finer nuances than just accepting or rejecting monotonicity.

This article proposes a Bayesian approach to evaluating manifest monotonicity. First, some hypotheses are considered that could be deemed relevant in the context of manifest monotonicity. These hypotheses require us to go beyond the standard null hypothesis testing framework. We argue that being able to distinguish between these different hypotheses can be highly relevant in the context of IRT. In the next section, we elaborate how one can compare these different hypotheses by using the Bayes factor to quantify the relative support for these hypotheses. Using a simulation study, the performance of the procedure is evaluated under varying conditions. Finally, the relevance and the application of this procedure is illustrated using an empirical example.

## 3.2 Relevant Competing Hypotheses

For a test with $k$ dichotomous items, let $X_i$ denote the score on item $i$, with realization $x_i = 0, 1$ for a negative and positive score, respectively. Let $\theta$ denote the latent variable. The assumption of latent monotonicity specifies that the item response function (IRF), $P(X_i = 1|\theta)$, is nondecreasing in $\theta$ (Hambleton & Swaminathan, 1985). A general formulation of the manifest

score, denoted by $Y$ and with realization $y$, can be obtained by using

$$Y = \sum_{i=1}^{k} c_i X_i, \tag{3.1}$$

where $c_1, \ldots, c_k$ are item inclusion coefficients that are chosen by the researcher, and $c_i \in \{0, 1\}$ for all $i$. The manifest score $Y$ is thus obtained by specifying particular items to be included or excluded using the item coefficients. For example, by setting $c_j = 0$ and $c_i = 1$ for all $i \neq j$, one obtains the unweighted restscore for item $j$. Although other manifest scores could be considered, the restscore is a more reliable ordinal estimator of the latent variable than a manifest score that is based on fewer items. Using the total score instead of the restscore would potentially confound the evaluation of latent monotonicity, since including the item in question in the manifest score might bias the results in the direction of monotonicity due to the correlation between the item and the total score (or potentially *against* monotonicity, see Junker & Sijtsma, 2000). For these reasons, the manifest score that is used in the remainder of the article is the unweighted restscore, although the proposed procedures can readily be extended to other manifest scores as well.

Let $h$ denote the highest possible value on manifest score $Y$. Regardless of which inclusion coefficients were selected, it can be obtained through $h = \sum_{i=1}^{k} c_i$. Furthermore, let $\pi_y = P(X = 1|Y = y)$ for the item that is investigated, where the subscript $i$ is dropped for notational convenience. The hypothesis that manifest monotonicity over $Y$ holds for a specific item corresponds to

$$H_{MM}: \quad \pi_0 \leq \ldots \leq \pi_y \leq \ldots \leq \pi_h.$$

$H_{MM}$ corresponds to the null hypothesis in the order-constrained statistical inference framework discussed by Tijmstra et al. (2013), and it can be contrasted with its negation, the claim that there are nonmonotonicities at the manifest level:

$$\begin{array}{lll} H_{NM}: & H_{MM} \text{ does not hold}, & \text{or equivalently}, \\ & \pi_y > \pi_{y+1}, & \text{for at least one value of } y. \end{array}$$

Since these hypotheses are exhaustive and exclude each other, evaluating manifest monotonicity effectively boils down to choosing between $H_{MM}$ and $H_{NM}$. However, $H_{NM}$ is quite general, and hence not very informative. That is, if one ends up accepting $H_{NM}$, there is very little that can be said about the ordering of the conditional item probabilities $\pi_0, \ldots, \pi_h$, other than the fact that their ordering is not completely monotone. Following the terminology of Hoijtink (2012), $H_{NM}$ can be considered to have a high complexity, or similarly, to be relatively unspecific or uninformative.

In practical applications, one could be interested in finding out to which extent manifest monotonicity holds, that is, whether the conditional item probabilities show an ordering that is very similar or very dissimilar to the order specified by manifest monotonicity. Even though manifest monotonicity may not hold completely, it could be the case that the ordering of the conditional item probabilities still approximates monotonicity to a high degree. Such items could be considered to be *essentially* monotone, implying that the ordering only shows small deviations from those specified by $H_{MM}$. For example, one could define essential monotonicity as a less restrictive version of manifest monotonicity as specified in $H_{MM}$, where essential monotonicity allows for local violations of manifest monotonicity ($\pi_y > \pi_{y+1}$ for some $y$) as long as these violations only occur between adjacent values of $Y$. Depending on the specific practical or theoretical context, one might or might not be satisfied with including essentially monotone items in a test, and hence finding out whether items are strictly monotone, essentially monotone, or not monotone at all can be of interest to for example test constructors. With these considerations in mind, it may be of use to evaluate the hypothesis that a form of 'essential monotonicity' holds for a specific item, which could be formulated as:

$$
\begin{aligned}
H_{EM}: \quad & \pi_0 \leq min\{\pi_2, \pi_3\}, \\
& \pi_1 \leq min\{\pi_3, \pi_4\}, \\
& \quad \vdots \\
& \pi_{h-3} \leq min\{\pi_{h-1}, \pi_h\}, \\
& \pi_{h-2} \leq \pi_h.
\end{aligned}
$$

This formulation of essential monotonicity shows that $H_{EM}$ allows for violations of manifest monotonicity ($\pi_a > \pi_b$ for some $a < b$), but only if these violations occur between two adjacent values of $Y$. That is, essential monotonicity is violated as soon as for some $y$, $\pi_y > \pi_{y+d}$ for some $d \in \{2, \ldots, h-y\}$. More liberal versions of essential monotonicity can be obtained by letting $d \in \{e, \ldots, h-y\}$, where $e > 2$. Of course, the more one increases $e$, the less restrictive and less informative $H_{EM}$ becomes, up to the point where monotonicity is hardly captured by $H_{EM}$ any more. A practical benefit of considering essential monotonicity, aside from the fact that it could be of substantial interest, is that it could help increase the power to detect small violations of manifest monotonicity (as will be elaborated in the next section). This potential increase in power originates from the fact that $H_{EM}$ places much more restrictions on the conditional item probabilities than $H_{NM}$, and hence is more specific.

Another alternative to $H_{MM}$ that one might want to consider is the postulation of a ceiling or a floor effect, formulated in $H_C$ and $H_F$ respectively:

$$\begin{aligned} H_C: &\quad \pi_0 \leq \ldots \leq \pi_{c-1} \leq \{\pi_c, \ldots, \pi_h\} \\ H_F: &\quad \{\pi_0, \ldots, \pi_f\} \leq \pi_{f+1} \leq \ldots \leq \pi_h, \end{aligned}$$

where $c$ denotes the 'ceiling-value' and $f$ the 'floor-value' of the manifest score. Both $H_C$ and $H_F$ leave the ordering of some of the conditional item probabilities open, thus allowing for nonmonotonicities above ($H_C$) or below ($H_F$) a certain value on the manifest score. This weaker form of monotonicity could be of interest for selection or testing purposes, if the main goal of a test is to distinguish between respondents on either the low or the high end of the distribution, but not necessarily over the entire range of the latent variable. Additionally, such hypotheses could be of substantial relevance in the context of exam items, where the possibility of providing the desired answer may actually decrease if the examinee progresses too far on the latent variable, or in the context of multiple choice items where certain 'distractor' qualities of wrong alternatives fail to function if the examinee's ability is too low.

Like $H_{EM}$, $H_C$ and $H_F$ are more specific than $H_{NM}$, which again could

result in increased power to detect violations of monotonicity. Focussing on these specific kind of deviations from monotonicity could thus result in an increase in power to detect these violations, and could also have substantial relevance in some realistic application settings of IRT. The section dealing with the empirical example is meant to illustrate the added value of considering such informed alternative hypotheses. To deal with these hypotheses, the use of Bayes factors first has to be discussed, which is the topic of the next section.

With these considerations in mind, it may be of use to evaluate the hypothesis that a form of 'essential monotonicity' holds for a specific item, which could be formulated as:

$$\begin{aligned} H_{EM}: \quad & \pi_0 \leq min\{\pi_2, \pi_3\}, \\ & \pi_1 \leq min\{\pi_3, \pi_4\}, \\ & \quad \vdots \\ & \pi_{h-3} \leq min\{\pi_{h-1}, \pi_h\}, \\ & \pi_{h-2} \leq \pi_h. \end{aligned}$$

This formulation of essential monotonicity shows that $H_{EM}$ allows for violations of manifest monotonicity ($\pi_a > \pi_b$ for some $a < b$), but only if these violations occur between two adjacent values of $Y$. That is, essential monotonicity is violated as soon as for some $y$, $\pi_y > \pi_{y+d}$ for some $d \in \{2, \ldots, h-y\}$. More liberal versions of essential monotonicity can be obtained by letting $d \in \{e, \ldots, h-y\}$, where $e > 2$. Of course, the more one increases $e$, the less restrictive and less informative $H_{EM}$ becomes, up to the point where monotonicity is hardly captured by $H_{EM}$ any more. A practical benefit of considering essential monotonicity, aside from the fact that it could be of substantial interest, is that it could help increase the power to detect small violations of manifest monotonicity (as will be elaborated in the next section). This potential increase in power originates from the fact that $H_{EM}$ places much more restrictions on the conditional item probabilities than $H_{NM}$, and hence is more specific.

Another alternative to $H_{MM}$ that one might want to consider is the postulation of a ceiling or a floor effect, formulated in $H_C$ and $H_F$ respec-

tively:

$$\begin{aligned} H_C: &\quad \pi_0 \leq \ldots \leq \pi_{c-1} \leq \{\pi_c, \ldots, \pi_h\} \\ H_F: &\quad \{\pi_0, \ldots, \pi_f\} \leq \pi_{f+1} \leq \ldots \leq \pi_h, \end{aligned}$$

where $c$ denotes the 'ceiling-value' and $f$ the 'floor-value' of the manifest score. Both $H_C$ and $H_F$ leave the ordering of some of the conditional item probabilities open, thus allowing for nonmonotonicities above ($H_C$) or below ($H_F$) a certain value on the manifest score. This weaker form of monotonicity could be of interest for selection or testing purposes, if the main goal of a test is to distinguish between respondents on either the low or the high end of the distribution, but not necessarily over the entire range of the latent variable. Additionally, such hypotheses could be of substantial relevance in the context of exam items, where the possibility of providing the desired answer may actually decrease if the examinee progresses too far on the latent variable, or in the context of multiple choice items where certain 'distractor' qualities of wrong alternatives fail to function if the examinee's ability is too low.

Like $H_{EM}$, $H_C$ and $H_F$ are more specific than $H_{NM}$, which again could result in increased power to detect violations of monotonicity. Focussing on these specific kind of deviations from monotonicity could thus result in an increase in power to detect these violations, and could also have substantial relevance in some realistic application settings of IRT. The section dealing with the empirical example is meant to illustrate the added value of considering such informed alternative hypotheses. To deal with these hypotheses, the use of Bayes factors first has to be discussed, which is the topic of the next section.

With these considerations in mind, it may be of use to evaluate the hypothesis that a form of 'essential monotonicity' holds for a specific item, which could be formulated as:

$$\begin{aligned} H_{EM}: &\quad \pi_0 \leq min\{\pi_2, \pi_3\}, \\ &\quad \pi_1 \leq min\{\pi_3, \pi_4\}, \\ &\qquad \vdots \\ &\quad \pi_{h-3} \leq min\{\pi_{h-1}, \pi_h\}, \\ &\quad \pi_{h-2} \leq \pi_h. \end{aligned}$$

This formulation of essential monotonicity shows that $H_{EM}$ allows for violations of manifest monotonicity ($\pi_a > \pi_b$ for some $a < b$), but only if these violations occur between two adjacent values of $Y$. That is, essential monotonicity is violated as soon as for some $y$, $\pi_y > \pi_{y+d}$ for some $d \in \{2, \ldots, h-y\}$. More liberal versions of essential monotonicity can be obtained by letting $d \in \{e, \ldots, h-y\}$, where $e > 2$. Of course, the more one increases $e$, the less restrictive and less informative $H_{EM}$ becomes, up to the point where monotonicity is hardly captured by $H_{EM}$ any more. A practical benefit of considering essential monotonicity, aside from the fact that it could be of substantial interest, is that it could help increase the power to detect small violations of manifest monotonicity (as will be elaborated in the next section). This potential increase in power originates from the fact that $H_{EM}$ places much more restrictions on the conditional item probabilities than $H_{NM}$, and hence is more specific.

Another alternative to $H_{MM}$ that one might want to consider is the postulation of a ceiling or a floor effect, formulated in $H_C$ and $H_F$ respectively:

$$\begin{aligned} H_C: &\quad \pi_0 \leq \ldots \leq \pi_{c-1} \leq \{\pi_c, \ldots, \pi_h\} \\ H_F: &\quad \{\pi_0, \ldots, \pi_f\} \leq \pi_{f+1} \leq \ldots \leq \pi_h, \end{aligned}$$

where $c$ denotes the 'ceiling-value' and $f$ the 'floor-value' of the manifest score. Both $H_C$ and $H_F$ leave the ordering of some of the conditional item probabilities open, thus allowing for nonmonotonicities above ($H_C$) or below ($H_F$) a certain value on the manifest score. This weaker form of monotonicity could be of interest for selection or testing purposes, if the main goal of a test is to distinguish between respondents on either the low or the high end of the distribution, but not necessarily over the entire range of the latent variable. Additionally, such hypotheses could be of substantial relevance in the context of exam items, where the possibility of providing the desired answer may actually decrease if the examinee progresses too far on the latent variable, or in the context of multiple choice items where certain 'distractor' qualities of wrong alternatives fail to function if the examinee's ability is too low.

Like $H_{EM}$, $H_C$ and $H_F$ are more specific than $H_{NM}$, which again could

result in increased power to detect violations of monotonicity. Focussing on these specific kind of deviations from monotonicity could thus result in an increase in power to detect these violations, and could also have substantial relevance in some realistic application settings of IRT. The section dealing with the empirical example is meant to illustrate the added value of considering such informed alternative hypotheses. To deal with these hypotheses, the use of Bayes factors first has to be discussed, which is the topic of the next section.

## 3.3 Bayes Factors

When two competing hypotheses are considered, their relative support can be quantified using the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). This measure balances the fit of the different hypotheses with their complexity. To determine the fit and the complexity of a hypothesis $H_Z$ imposing order constraints on $\pi_0, \ldots, \pi_h$, a prior distribution of $\boldsymbol{\pi}$ needs to be specified, and the posterior distribution of $\boldsymbol{\pi}$ after observing the data needs to be determined.

In order to ensure that every ordering of $\pi_0, \ldots, \pi_h$ is equally likely a priori, one can specify the prior distribution to be

$$h(\boldsymbol{\pi}) = \prod_{y=0}^{h} Beta(\pi_y|1,1) = 1. \tag{3.2}$$

This prior distribution does not favor any specific ordering of $\pi_0, \ldots, \pi_h$, and for each $\pi_y$ assigns equal probability to all values between 0 and 1. For this reason it can be considered to be uninformative (Lynch, 2007). Using the prior distribution that is specified in Equation 3.2, the complexity of every inequality constrained hypothesis can be determined analytically.

Assuming the responses to the item in question to be binomially distributed for each level of the manifest score, the likelihood of the data corresponds to

$$f(\mathbf{X}|\boldsymbol{\pi}) = \prod_{y=0}^{h} \pi_y^{s_y}(1-\pi_y)^{n_y-s_y}, \tag{3.3}$$

where $\mathbf{X}$ denotes the vector containing the responses to the item in question, $n_y$ denotes the number of respondents with manifest score $y$, and $s_y$ denotes the number of respondents with manifest score $y$ that received a score of 1 on the item. The posterior distribution of the conditional item probabilities is proportional to the product of the likelihood and the prior distribution, and corresponds to

$$g(\boldsymbol{\pi}|\mathbf{X}) = \prod_{y=0}^{h} Beta(\pi_y|s_y + 1, n_y - s_y + 1). \tag{3.4}$$

In a Bayesian framework, the complexity $c_Z$ of a hypothesis $H_Z$ can be defined as the proportion of the prior distribution of $\boldsymbol{\pi}$ that is in accordance with that hypothesis (Hoijtink, 2012). Thus, for a hypothesis $H_Z$,

$$c_Z = \frac{\int h(\boldsymbol{\pi}) I_{\boldsymbol{\pi} \in H_Z} d\boldsymbol{\pi}}{\int h(\boldsymbol{\pi}) d\boldsymbol{\pi}}, \tag{3.5}$$

where $I_{\boldsymbol{\pi} \in H_Z}$ is an indicator function that takes on a value of 1 if the values in $\boldsymbol{\pi}$ are in correspondence with $H_Z$, and 0 otherwise. Thus, the complexity of a hypothesis such as $H_{MM}$ corresponds to the probability of obtaining a set of values for $\boldsymbol{\pi}$ that match the constraints specified by $H_{MM}$ if we were to randomly draw values from the prior distribution of $\boldsymbol{\pi}$.

In a similar vein, the posterior fit $f_Z$ of hypothesis $H_Z$ to the data can be defined as the proportion of the posterior distribution of $\boldsymbol{\pi}$ that is in accordance with that hypothesis (Hoijtink, 2012), corresponding to

$$f_Z = \frac{\int g(\boldsymbol{\pi}|\mathbf{X}) I_{\boldsymbol{\pi} \in H_Z} d\boldsymbol{\pi}}{\int g(\boldsymbol{\pi}|\mathbf{X}) d\boldsymbol{\pi}}. \tag{3.6}$$

By comparing the fit of a hypothesis with its complexity, one can determine the extent to which the data provide evidence in favor or against that hypothesis. The ratio $\frac{f}{c}$ reflects the amount of support that the hypothesis receives from the data, since it quantifies how much more likely the hypothesis has become after observing the data (Kass & Raftery, 1995). The Bayes factor comparing two competing hypotheses that specify order

constraints for $\boldsymbol{\pi}$ can be calculated simply by taking the ratio of $\frac{f}{c}$ of the two hypotheses (Hoijtink, 2012). Thus, the Bayes factor indicates which hypothesis receives more support from the data. It should be noted that the Bayes factor does *not* indicate which of the two hypotheses is more likely to be true, since it does not simply compare the posterior fit of the two hypotheses but rather the extent to which the hypotheses have become more likely after observing the data. In this way, the Bayes factor of a highly specific hypothesis (i.e., with a small $c$) that still has a good fit (i.e., high $f$) will be higher than that of an unspecific hypothesis (i.e., high $c$) that has the same fit, rewarding the more specific hypothesis for its parsimony.

With regard to manifest monotonicity, the simplest comparison that can be made is between $H_{MM}$ and the unconstrained alternative $H_U$ : $\{\pi_0, \ldots, \pi_h\}$. The corresponding Bayes factor can be computed through

$$BF_{MM,U} = \frac{\frac{f_{MM}}{c_{MM}}}{\frac{f_U}{c_U}} = \frac{f_{MM}}{c_{MM}}.$$

Here, $\frac{f_U}{c_U}$ drops out of the equation, since $H_U$ does not restrict $\boldsymbol{\pi}$ and hence $f_U = c_U$. A value of $BF_{MM,U}$ greater than 1 indicates that the data provide support for $H_{MM}$, while a value smaller than 1 indicates that the data do not support the hypothesis of manifest monotonicity.

Contrasting $H_{MM}$ with $H_U$ is not very informative, since $H_U$ incorporates $H_{MM}$ and hence does not contradict it. In order to evaluate $H_{MM}$, this hypothesis should be contrasted with a competing hypothesis. For example, $H_{MM}$ can be contrasted with its complement, $H_{NM}$:

$$BF_{MM,NM} = \frac{f_{MM}c_{NM}}{f_{NM}c_{MM}} = \frac{f_{MM}(1 - c_{MM})}{(1 - f_{MM})c_{MM}}.$$

Thus, $BF_{MM,NM}$ quantifies the amount of support that $H_{MM}$ receives from the data when contrasted with its complement.

The comparison of $H_{MM}$ and $H_{NM}$ provides useful information about the general support for the hypothesis that the conditional item probabilities are ordered in accordance with manifest monotonicity. Thus, the

comparison provides a direct and useful quantification of the amount of support the data provide for this property. However, contrasting $H_{MM}$ with one of the more specific alternative hypotheses holds some benefits as well. First, if one ends up accepting $H_{NM}$ over $H_{MM}$, there is little that can be said about the ordering of $\pi_0, \ldots, \pi_h$, since $H_{NM}$ is simply defined as the negation of $H_{MM}$ and contains all $(h+1)!$ possible orderings excluding the one allowed by $H_{MM}$. In practice, finding out the extent and the location of the violations of monotonicity can be of interest, and using more informative alternative hypotheses like the ones discussed in the previous section may help achieve this goal. Additionally, precisely because $H_{NM}$ is highly unspecific, it also includes many possible orderings that are far removed from the true ordering. If manifest monotonicity is only violated locally, many of the alternative orderings that are included in $H_{NM}$ will have worse fit to the data than the ordering prescribed by $H_{MM}$. Since in calculating $f_{NM}$ one integrates over the posterior distribution for the entire area not included in $H_{MM}$, these unlikely orderings included in $H_{NM}$ reduce the overall support for $H_{NM}$. The inclusion of unlikely orderings in $H_{NM}$ detracts from the sensitivity to detect violations of manifest monotonicity.

By only considering a subset of the orderings that $H_{NM}$ allows, manifest monotonicity can be contrasted with more specific alternatives. If realistic alternatives are selected, the power to detect violations of manifest monotonicity can be increased, since these alternatives may receive more support from the data than the uninformative $H_{NM}$. For example, one might want to consider contrasting $H_{MM}$ with $H_{EM}$, thereby excluding all orderings that deviate from monotonicity by a large extent. Such an approach can be particularly useful in the context of already existing tests where the items have been specifically designed to measure a common variable, and where it is reasonable to assume that deviations from monotonicity are at most small ones. In order to make the hypotheses exhaustive, one can define $H_{EM'}$ as $H_{EM}$ with the added constraint that $H_{MM}$ does not hold. For

this comparison, one obtains

$$BF_{MM,EM'} = \frac{f_{MM}c_{EM'}}{f_{EM'}c_{MM}} = \frac{f_{MM}(c_{EM} - c_{MM})}{(f_{EM} - f_{MM})c_{MM}}.$$

Similarly, one can contrast $H_{MM}$ with $H_{C'}$ or $H_{F'}$, where $H_{C'}$ or $H_{F'}$ are obtained from $H_C$ and $H_F$ by adding the constraint that $H_{MM}$ does not hold. This way, one can evaluate $BF_{MM,C'}$ and $BF_{MM,F'}$, which indicate whether there is a reason to suspect that monotonicity is violated at the high end or the low end of the manifest score respectively. Furthermore, if for a particular application all that is relevant is that monotonicity holds either for the low or for the high end of the manifest score, this can be evaluated by contrasting $H_C$ or $H_F$ with their complements rather than making use of $H_{MM}$. This allows for a lot of flexibility in evaluating the precise property that is relevant in a particular situation.

Regardless of which specific hypotheses one wants to compare, the calculation of the Bayes factor requires one to obtain the fit and the complexity of the two hypotheses of interest. Under the uninformative prior distribution of $\boldsymbol{\pi}$ in Equation 3.2, each ordering of the conditional item probabilities is equally likely, and the complexity of any hypothesis $H_Z$ about the ordering of these conditional item probabilities can be obtained through

$$c_{Z,h} = \frac{O_{Z,h}}{(h+1)!},$$

where $O_{Z,h}$ denotes the number of possible orderings of the conditional item probabilities that are allowed by $H_Z$, given that the highest possible value on the manifest score is $h$. Hence, $O_{MM,h}$ is equal to 1, and $O_{C,h}$ and $O_{F,h}$ are equal to $(h-(c-1))!$ and $(f+1)!$, respectively. The number of orderings that essential monotonicity allows will always be a number from the Fibonacci sequence. That is, $O_{EM,h} = Fib(h+3)$, where $Fib = 0, 1, 1, 2, 3, 5, 8, 13, \ldots$. Because the constraints in $H_{EM}$ specify that conditional probabilities two scores apart cannot decrease, increasing $h$ by 1 increases the number of acceptable orderings by $O_{EM,h-1}$. That is, when $h$ increases by 1, the highest possible manifest score becomes $h+1$, and

there are two types of orderings possible that are allowed by $H_{EM}$: orderings where $\pi_h \leq \pi_{h+1}$, of which there are $O_{EM,h}$ in total, and orderings where $\pi_{h-1} \leq \pi_{h+1} < \pi_h$, of which there are $O_{EM,h-1}$. Thus, for any $h > 0$, $O_{EM,h+1} = O_{EM,h} + O_{EM,h-1}$, resulting in the Fibonacci sequence. The complexity of $H_{EM'}$, $H_{C'}$ and $H_{F'}$ can be obtained by subtracting 1 from $O_{EM,h}$, $O_{C,h}$ and $O_{F,h}$, respectively.

To determine the fit of the different hypotheses, the posterior distribution of $\boldsymbol{\pi}$ specified in Equation 3.4 needs to be evaluated. While the posterior distribution of $\boldsymbol{\pi}$ is proportional to the product of its prior distribution and the likelihood of the data, analytically determining the fit of the hypotheses is not straightforward. Instead of exact integration, a Gibbs sampling procedure can be used to approximate the proportion of the posterior that falls within the specified part of the parameter space. This procedure enables one to repeatedly sample values of $\boldsymbol{\pi}$ from its posterior distribution, thus allowing one to approximate the posterior distribution to any degree of precision and hence making it possible to approximate the value of $f_Z$ for any $H_Z$. However, since $f_Z$ could be extremely small for large values of $h$, estimating $f_Z$ simply by counting the proportion of draws from the posterior distribution of $\boldsymbol{\pi}$ that are in accordance with the constraints specified in $H_Z$ does not necessarily result in an accurate estimate of $f_Z$, unless one is willing to evaluate an enormously large number of draws.

A more elegant and computationally less demanding approach is to sequentially evaluate the individual constraints specified in $H_Z$. This can be done using a decomposition of the Bayes factor as discussed by Mulder, Klugkist, Van de Schoot, Meeus, Selfhout and Hoijtink (2009). In this decomposition, the Bayes factor of a hypothesis $H_Z$ with $z$ constraints against $H_U$ is decomposed into $m$ Bayes factors in the following way:

$$\begin{aligned} BF_{Z,U} &= BF_{1,U} * BF_{2,1} * \ldots * BF_{k,k-1} * \ldots * BF_{z,z-1} \\ &= \frac{f_{1|U}}{c_{1|U}} * \frac{f_{2|1}}{c_{2|1}} * \ldots * \frac{f_{k|k-1}}{c_{k|k-1}} * \ldots * \frac{f_{z|z-1}}{c_{z|z-1}} \\ &= \frac{f_{1|U} * f_{2|1} * \ldots * f_{k|k-1} * \ldots * f_{z|z-1}}{c_Z} \end{aligned} \tag{3.7}$$

Here, $BF_{1,U}$ is the Bayes factor comparing the hypothesis that the first

order constraint holds ($H_1$) with the unconstrained hypothesis ($H_U$), and $BF_{k,k-1}$ denotes the Bayes factor which compares the hypothesis that the first $k$ order constraints hold ($H_k$) with the hypothesis that the first $k-1$ constraints hold ($H_{k-1}$). Furthermore, $f_{k|k-1}$ is the fit of $H_k$, conditional on the assumption that $H_{k-1}$ holds. For each hypothesis $H_k$, this conditional fit measure $f_{k|k-1}$ can be estimated using a Gibbs sampler that draws values from the posterior distribution under the $k-1$ constraints of $H_{k-1}$, that is

$$g(\boldsymbol{\pi}|\mathbf{X}; \boldsymbol{\pi} \in H_{k-1}) \propto \prod_{y=0}^{h} Beta(\pi_y|s_y+1, n_y-s_y+1) I_{\boldsymbol{\pi}\in H_{k-1}}. \qquad (3.8)$$

Thus, the full conditional posterior distribution of each $\pi_y$ is either a truncated beta distribution if $\pi_y$ is constrained by $H_{k-1}$, or a regular beta distribution otherwise. After allowing for a burn-in period (e.g., after discarding the first 5000 draws), these draws result in an approximation of the posterior distribution $g(\boldsymbol{\pi}|\mathbf{X}; \boldsymbol{\pi} \in H_{k-1})$ that can be used to estimate $f_{k|k-1}$ (e.g., using 10,000 draws). By sequentially applying this Gibbs sampler to estimate $f_{1|u}, \ldots, f_{z|z-1}$, $f_Z$ can be approximated. This procedure makes it possible to approximate the fit of any hypothesis imposing order constraints on $\boldsymbol{\pi}$.

This way, the Bayes factor can be obtained for any pair of order-constrained hypotheses about the conditional item probabilities. This procedure has been implemented as a function in R that can be used to evaluate manifest monotonicity, by contrasting $H_{MM}$ with $H_{NM}$ as well as $H_{EM'}$. The test function is available on request from the first author. After obtaining the Bayes factors, conclusions can be drawn about the extent to which the data support manifest monotonicity over its alternatives. Kass and Raftery (1995) provide some general guidelines about the interpretation of Bayes factors (extending and revising the suggestions made by Jeffreys, 1961). Their guidelines will be adopted in this paper. They argue that Bayes factors between $\frac{1}{3}$ and 3 provide little support one way or the other, and that only outside of that range Bayes factors begin to show support that is worth mentioning. A Bayes factor between 3 and 20 shows some support in favor of the first hypothesis, and a Bayes factor in the range of

$\frac{1}{20}$ through $\frac{1}{3}$ shows some support in favor of the second hypothesis. Bayes factors larger than 20 indicate strong support for the first hypothesis, and Bayes factors smaller than $\frac{1}{20}$ show a strong support for the second hypothesis. These are just rough-and-ready guidelines for a measure that is actually continuous, so the precise value of the Bayes factor is more important than simply categorizing it as 'strong support'. Still, practical applications will require the implementation of decision rules as to whether the observed Bayes factor is high enough to justify the assumption of latent monotonicity, and for this purpose specifying a 'critical value' for the Bayes factor is unavoidable. As a suggestion, one might consider accepting latent monotonicity if there is strong support for $H_{MM}$ over $H_{NM}$ ($BF > 20$) and at least some support for $H_{MM}$ over $H_{EM'}$ ($BF > 3$). Of course, any set of guidelines will be open to debate to some extent, but these guidelines may prove helpful when applying the procedure.

## 3.4 Simulation Study

### 3.4.1 Method

To assess the performance of the proposed procedure under varying conditions, a simulation study was performed. The procedure was used to assess manifest monotonicity for three different items: an item with a monotone IRF, an item with a flat IRF, and an item with a locally nonmonotone IRF. The monotone IRF corresponded to a two-parameter logistic IRF with a difficulty of 0 and a discrimination of 1. The item with the flat IRF was included to assess the performance of the procedure in the boundary case – no violation of monotonicity, but also no increase in the conditional item probabilities – and was obtained by specifying the item probability to be .5 for every value on the latent variable. The locally nonmonotone IRF was obtained using a simple polynomial extension of the two-parameter logistic model,

$$P(X_i = 1|\theta) = \frac{\exp\left(\alpha_{1i}(\theta - \beta_{1i}) + \alpha_{2i}(\theta - \beta_{2i})^2 + \alpha_{3i}(\theta - \beta_{3i})^3\right)}{1 + \exp\left(\alpha_{1i}(\theta - \beta_{1i}) + \alpha_{2i}(\theta - \beta_{2i})^2 + \alpha_{3i}(\theta - \beta_{3i})^3\right)},$$

where $\beta_{1i}$, $\beta_{2i}$ and $\beta_{3i}$ function as the item's location parameters for the different-order polynomials, and $\alpha_{1i}$, $\alpha_{2i}$ and $\alpha_{3i}$ determine the slope of the IRF (see also Tijmstra et al., 2013). To obtain an item with a moderate violation of latent monotonicity, $\alpha_{1i}$, $\alpha_{2i}$ and $\alpha_{3i}$ were set to 1, 1.2 and .25 respectively, and $\beta_{1i}$, $\beta_{2i}$ and $\beta_{3i}$ were set to 2.5, 1.6 and 1.5 respectively. The shapes of the three IRFs are displayed in Figure 3.1.

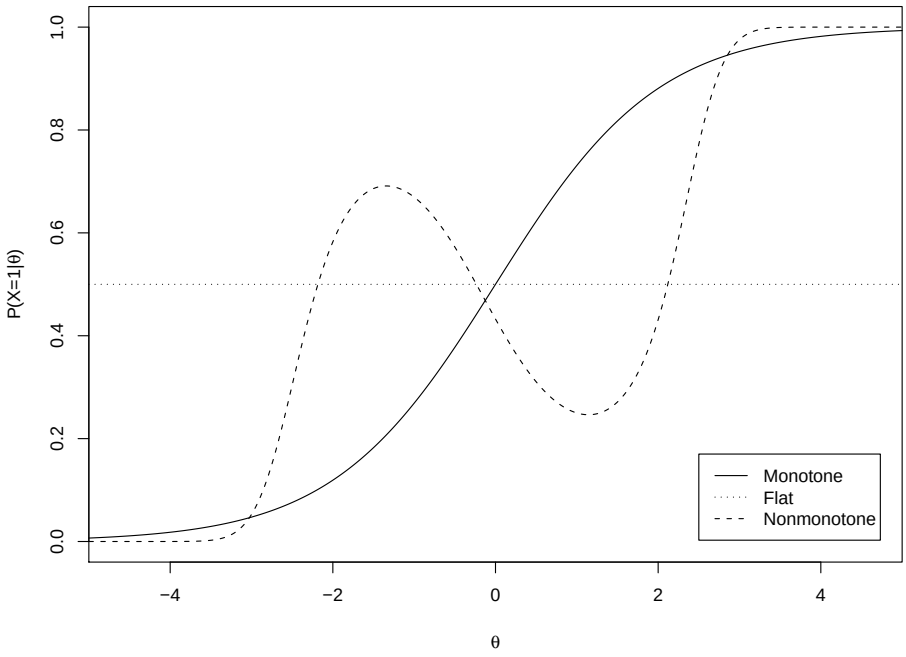


Figure 3.1: The item response functions of the three items that were analyzed. The monotone function is denoted by 'Monotone', the constant function is denoted by 'Flat', and the locally nonmonotone function is denoted by 'Nonmonotone'.

To assess manifest monotonicity for these items, manifest scores needed to be created. Three manifest scores were constructed: a restscore containing 5, 10 and 20 monotone items respectively. These were obtained by using a set of five two-parameter logistic items, with difficulty parameters ranging from -1.0 to 1.0 and discrimination parameters ranging from .50 to 1.50. For restscores containing 10 and 20 items, these five IRFs were simply used two and four times, respectively. The shape of the five IRFs is shown in Figure 3.2.

Different sample sizes were used to study the effect sample size had on the values of the Bayes factors and the resulting decision about manifest monotonicity that would have been made if the proposed guidelines would have been used. Sample sizes of 100, 200, 500 and 1000 respondents were used.

For each of the 3 by 3 by 4 conditions, 1000 replications were performed. For each replication, values on the latent variable were drawn from a standard normal distribution for each respondent, and subsequently data were generated for the item under consideration (monotone, non-monotone or flat) and the items that constitute the restscore (5, 10 or 20). Next, the Bayesian procedure was applied to the generated data, using 5000 iterations for the burn-in period of the Gibbs sampler and the subsequent 10,000 iterations to approximate the conditional posterior distribution $g(\boldsymbol{\pi}|\mathbf{X};\boldsymbol{\pi} \in H_{k-1})$ for each order constraint $k$, as detailed in Equation 3.7. This way, the Bayes factor of $H_{MM}$ versus $H_{NM}$ and $H_{MM}$ versus $H_{EM'}$ were obtained, and this was done for each of the 1000 sets of data to approximate the distribution of the two Bayes factors for that condition.

### 3.4.2 Results

Table 3.1 displays the results for the first part of the procedure, where $H_{MM}$ was contrasted with $H_{NM}$. For the monotone item, the proportion of replications where the Bayes factor indicated strong support for manifest monotonicity exceeded .80 for all conditions, excluding the condition where the sample size was 100 and the restscore was based on 5 items. The

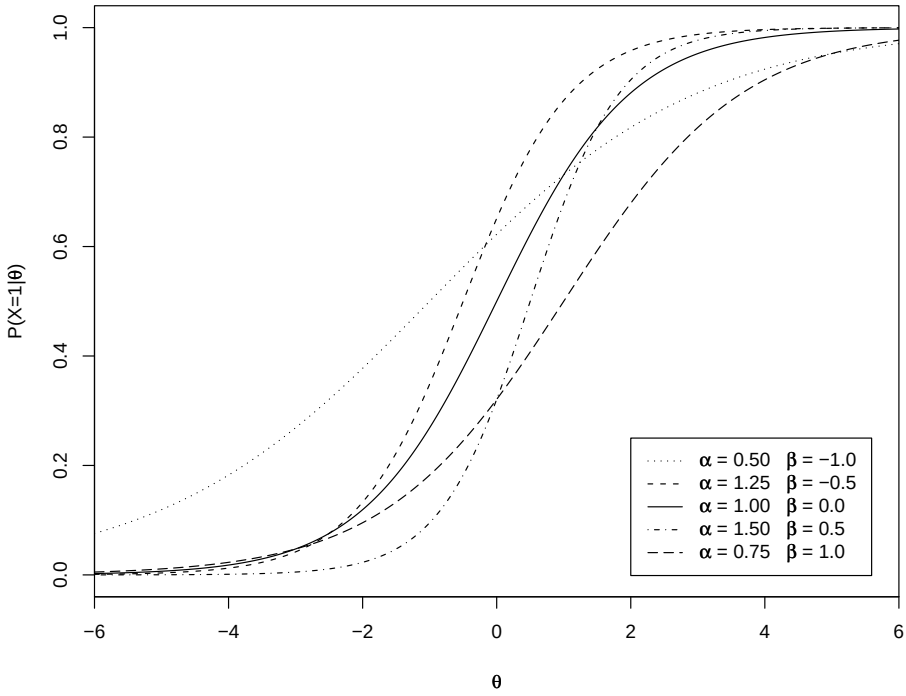


Figure 3.2: The item response functions of the five monotone items, based on the two-parameter logistic model. The discrimination and difficulty parameters are denoted by $\alpha$ and $\beta$, respectively.

proportion of cases with strong support *against* manifest monotonicity was always close to zero. As test length and sample size increased the proportion of replications with support for manifest monotonicity approached 1, indicating that with a sufficient amount of data the procedure correctly indicates that there is strong support for manifest monotonicity. Thus, in all but the most unfavorable conditions the monotone item had a very high

Table 3.1: Proportion of cases with strong support in favor of and against $H_{MM}$ when contrasted with $H_{NM}$ for the items with a monotone and a flat IRF (1,000 replications).

| | Strong support for $H_{MM}$ over $H_{NM}$ | | | | | |
|---|---|---|---|---|---|---|
| | Monotone IRF | | | Flat IRF | | |
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .583 | .841 | .936 | .018 | .082 | .094 |
| 200 | .858 | .965 | .995 | .016 | .102 | .145 |
| 500 | .981 | .998 | 1.000 | .024 | .097 | .206 |
| 1000 | .996 | 1.000 | .999 | .021 | .119 | .242 |

| | Strong support for $H_{NM}$ over $H_{MM}$ | | | | | |
|---|---|---|---|---|---|---|
| | Monotone IRF | | | Flat IRF | | |
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .003 | .005 | .004 | .197 | .262 | .378 |
| 200 | .000 | .000 | .002 | .174 | .320 | .345 |
| 500 | .000 | .000 | .000 | .171 | .300 | .331 |
| 1000 | .000 | .000 | .001 | .218 | .277 | .330 |

probability of correctly passing the first test of the procedure by receiving a Bayes factor larger than 20.

For the item with the flat IRF, it was expected that the proportion of replications with strong support one way or the other would be low. For this item, the proportion of replications with strong support for manifest monotonicity was indeed much lower than for the monotone item. Under most conditions the flat item had a high probability of not passing the first test of the procedure. However, the proportion of cases where there was strong support against manifest monotonicity was somewhat larger. The proportion of cases with strong support either in favor of or against manifest

monotonicity showed an increase as the test length increased. When the distribution of the Bayes factor is considered under the varying conditions (not displayed), it seems that the variance of the Bayes factor increases for this item when the test size becomes larger and to a lesser extent when the sample size increases. This may be taken as an indication that in these situations, the guideline of concluding that there is strong support for manifest monotonicity when the Bayes factor exceeds 20 is possibly too liberal, and that for large tests more extreme values for the Bayes factor are needed before one concludes that there is strong support one way or the other.

For reasons of conciseness the results for the item with a locally nonmonotone IRF are not displayed in Table 3.1, since the proportion of cases providing strong support for manifest monotonicity was always .001 or less. Thus, the locally nonmonotone item practically always fails the first test of the procedure, and continuing to the next step is not necessary for this item. Additionally, when the sample size was at least 200, the proportion of cases with strong support against manifest monotonicity exceeded .95 for this item. Thus, the procedure in its first step very easily picks up on the fact that the item is not completely monotone, illustrating that the procedure in this example has a very high power of detecting violations of monotonicity.

The second part of the procedure contrasts $H_{MM}$ with $H_{EM'}$. Since it is more difficult to distinguish between $H_{MM}$ and $H_{EM'}$, the proposed guidelines suggest that it is useful to focus on the cases where there is some support in favor of one of the hypotheses ($BF < \frac{1}{3}$ or $BF > 3$) rather than strong support. As can be observed in Table 3.2, for the monotone item the proportion of cases with support for manifest monotonicity over essential monotonicity varies greatly depending on test length and sample size. The proportion of cases where $H_{MM}$ was correctly supported over $H_{EM'}$ increased strongly as the sample size increased. A larger test size appears to make it more difficult to distinguish between the two hypotheses, as can be observed from the lower proportion of cases with support for $H_{MM}$. The explanation for this result may be that as the number of items included in the manifest score increases, the orderings specified by $H_{MM}$

Table 3.2: Proportion of cases with support in favor of and against $H_{MM}$ when contrasted with $H_{EM'}$ for the items with a monotone and a flat IRF (1,000 replications).

| | Support for $H_{MM}$ over $H_{EM'}$ | | | | | |
|---|---|---|---|---|---|---|
| | Monotone IRF | | | Flat IRF | | |
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .223 | .000 | .015 | .002 | .007 | .094 |
| 200 | .495 | .104 | .032 | .002 | .006 | .147 |
| 500 | .819 | .489 | .234 | .004 | .014 | .270 |
| 1000 | .953 | .815 | .519 | .003 | .010 | .362 |

| | Support for $H_{EM'}$ over $H_{MM}$ | | | | | |
|---|---|---|---|---|---|---|
| | Monotone IRF | | | Flat IRF | | |
| | $k$ | | | $k$ | | |
| $n$ | 5 | 10 | 20 | 5 | 10 | 20 |
| 100 | .039 | .029 | .006 | .047 | .036 | .054 |
| 200 | .026 | .042 | .003 | .042 | .038 | .074 |
| 500 | .003 | .022 | .013 | .050 | .046 | .084 |
| 1000 | .003 | .008 | .020 | .041 | .041 | .086 |

and $H_{EM'}$ become more and more similar. Thus, the larger the test length, the larger the sample size has to be to have a high probability of finding support for manifest monotonicity over its informed alternative.

The results for the flat item can also be found in Table 3.2. When 5 or 10 items are included in the manifest score, only about 1 percent of replications result in support for manifest monotonicity over essential monotonicity. Thus, for short tests contrasting $H_{MM}$ with $H_{EM'}$ forms an effective way of filtering out items that do not show a monotone increase in the conditional item probabilities. However, for longer tests the proportion of cases with support for manifest monotonicity increases. The reason for this seems to

Table 3.3: Conditional proportions $p_y$ and Bayes factors for the eight Raven scale items.

| Item | $p_0$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $BF_{MM,NM}$ | $BF_{MM,EM'}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .14 | .00 | .03 | .07 | .12 | .11 | .15 | .24 | 36.6 | 1.51 |
| 2 | .25 | .15 | .12 | .14 | .22 | .39 | .46 | .63 | 147.4 | 5.28 |
| 3 | .14 | .08 | .21 | .22 | .37 | .51 | .46 | .56 | 593.6 | 2.21 |
| 4 | .00 | .14 | .28 | .36 | .45 | .62 | .71 | .71 | 7656.1 | 13.15 |
| 5 | .00 | .19 | .35 | .54 | .73 | .75 | .81 | .71 | 1814.6 | 8.00 |
| 6 | .45 | .44 | .61 | .68 | .68 | .80 | .84 | 1.00 | 2637.0 | 4.82 |
| 7 | .45 | .67 | .82 | .78 | .80 | .83 | .87 | 1.00 | 462.4 | 3.97 |
| 8 | .65 | .74 | .84 | .90 | .93 | .96 | 1.00 | 1.00 | 927.4 | 4.62 |

be that the Bayes factor for the flat item again shows larger variance when test length increases. Thus, these results indicate that contrasting $H_{MM}$ with $H_{EM'}$ forms an effective way of filtering out items that do not show a monotone increase in the conditional item probabilities, but only for tests that are not too long.

## 3.5 Empirical Example

The procedure outlined above was applied to evaluate manifest monotonicity for each item from a set of eight Raven matrices items (Raven, 1956). These items required respondents to use inductive reasoning to determine what kind of figure is needed to correctly complete a three by three matrix of figures, and are often used in intelligence tests. Responses to the eight items were obtained from 494 freshman psychology students, with correct answers coded as '1' and other responses as '0'. For each of the items, the Bayes factor contrasting the support for $H_{MM}$ with that for $H_{NM}$ was determined, as well as the Bayes factor contrasting $H_{MM}$ with $H_{EM'}$. After discarding the first 5,000 draws for the burn-in period, 10,000 draws from the posterior distribution were used to evaluate each order constraint. The results of the analysis are displayed in Table 3.3.

For the comparison of manifest monotonicity with its complement, the values of $BF_{MM,NM}$ ranged from 36.6 to 7656.1, indicating that the data provide strong support for manifest monotonicity over its complement for each of the items. Thus, for each of the items there seems to be a general positive trend in the conditional item probabilities. The observed variance in the values of the Bayes factor seems to be partly attributable to the variance of these conditional item probabilities, with items for which these probabilities increase more strongly over the restscore receiving a much larger Bayes factor than items for which this overall increase was smaller.

While these values of $BF_{MM,NM}$ suggest that there is general support for latent monotonicity, one would ideally like to exclude the possibility that the item response functions are only largely monotone. For example, for the most difficult item (item 1) the proportion of correct answers in the highest restscore group was only .24, and the respondents in the lowest restscore group scored better than the respondents in the next five restscore groups. This local deviation could be due to sampling error, but could also reflect a genuine violation of monotonicity. Since most of the other proportions did show a monotone ordering, this potential violation might remain masked if one only considers $BF_{MM,NM}$. Although with a value of 35.6 item 1 had the lowest Bayes factor of the set, the Bayes factor stills indicate that there is strong support for manifest monotonicity for this item. To evaluate whether it is indeed plausible that manifest monotonicity holds for this item and for the other items, additionally contrasting manifest monotonicity with more informative alternatives is useful.

To more thoroughly evaluate the support for monotonicity, $H_{MM}$ was also contrasted with its close alternative, $H_{EM'}$. Since essential monotonicity could be a plausible hypothesis for items such as these that have been constructed to measure a common ability and that will likely show a positive correlation, using essential monotonicity as an alternative hypothesis to contrast manifest monotonicity with can be useful to determine if an item is not just generally monotone but strictly monotone. To eliminate potential worries about local violations one would like to find at least some support for $H_{MM}$ over $H_{EM}$. For these eight items, $BF_{MM,EM'}$ ranged from 1.51 to 13.15. Although for all items there was more support for $H_{MM}$ than for

$H_{EM'}$, none of the items showed strong support for manifest monotonicity over essential monotonicity. Six items showed some support for manifest monotonicity ($3 < BF_{MM,EM'} < 20$), and the lowest Bayes factor was obtained for the item that also received the lowest value on $BF_{MM,NM}$ (item 1).

With the suggested decision rules in mind, these results indicate that there is strong general support for manifest monotonicity for all eight items, thus supporting the idea that the assumption of latent monotonicity holds for these items. Six items also showed support for manifest monotonicity over its informed alternative, essential monotonicity, further supporting the idea that latent monotonicity holds for these items. Two items did not show support for manifest monotonicity over essential monotonicity, and hence this step of the procedure did not provide additional support for latent monotonicity for these two items. While the support for manifest monotonicity over its complement indicates that applying an item response model to these items might be appropriate, researchers would be well advised to have a close and critical look at the way these items function.

## 3.6 Discussion

The evaluation of model assumptions such as latent monotonicity plays a crucial part in the practical application of statistical models to real life data. Whereas existing approaches to testing manifest monotonicity have focussed on falsifying this assumption, the current approach provides an alternative methodology that focuses on evaluating the amount of support the data provide in favor of manifest monotonicity. This focus on confirmation instead of falsification is in line with the idea that a failure to reject a null hypothesis does not confirm that hypothesis – it is only retained. While a falsification-based approach might suit the critical evaluation of substantive hypotheses, model assumptions seem to demand for a more confirmatory approach like the one presented in this article. That is, before applying a certain model, we do not simply want to have failed to falsify its assumptions. Instead, one needs support for the validity of the assumptions for

the application that one has in mind. The proposed approach provides such a procedure, by using the Bayes factor to summarize the amount of support that manifest monotonicity receives from the data compared to an alternative where monotonicity does not hold. The procedure itself is neutral with regard to the decision to aim at verification or falsification, since it simply quantifies the support for or against manifest monotonicity. If the aim is falsification rather than verification, one can simply change the proposed guidelines and decide to reject latent monotonicity if there is a certain amount of support against manifest monotonicity.

By determining the relative support for manifest monotonicity as compared to its complement, the procedure provides a general measure of the amount of support for this property. Since this complement is very unspecific, the procedure can be made more powerful by subsequently comparing manifest monotonicity with an informed alternative hypothesis. These informed alternatives can either serve as substantive alternatives with a meaningful interpretation (such as the discussed floor and ceiling effects), or simply as way of more extensively investigating the amount of support in favor of manifest monotonicity (such as essential monotonicity). Since the Bayes factor can be determined for any set of order constraints on the conditional item probabilities, this approach is very flexible with regard to the range of hypotheses that can be compared. The procedure can also readily be extended to assess monotonicity for a set of items at once, since this can be achieved by simply multiplying the individual Bayes factors. Such an approach would however run the risk of masking violations for a particular item if the other items are monotone, so it seems that analysis at the item level is to be preferred.

The simulation results show that the first step of the procedure performs well under the null situation. That is, when a monotone item was considered, the procedure showed a high percentage of correct decisions when comparing manifest monotonicity with its complement in almost all conditions. This percentage increases as the number of persons that take the test or the number of items on the test increases. This first step was already sufficient to detect the violation of monotonicity present in the non-monotone item that was considered. However, the first step occasionally

failed to filter out the item with a flat IRF. Including a second step in the procedure where manifest monotonicity was contrasted with essential monotonicity turned out to be useful, as it helped to filter out items like this.

The results for the second part of the procedure – where manifest monotonicity is contrasted with essential monotonicity – showed that distinguishing these two types of orderings is more difficult. That is, for the monotone item support for manifest monotonicity over essential monotonicity was only found frequently when sample size was large. Longer tests also seem to require larger sample sizes before these two orderings can be distinguished sufficiently. This could be an indication that for long tests it is useful to employ a more liberal version of essential monotonicity in order to successfully differentiate between a completely monotone ordering and largely monotone orderings of the conditional item probabilities. Additionally, these results illustrate that not finding support for manifest monotonicity over essential monotonicity does not in itself imply that manifest monotonicity is violated.

It should be emphasized that the Bayes factor provides a measure of *relative* support. As such, it does not directly inform the researcher of the probability that the assumption is true, but rather the extent to which this has become more likely after observing the data. As such, the Bayes factor avoids leaning too heavily on the specification of the prior probability of the assumption being true. That is, while one could also opt to pursue a similar approach to obtain the probability of the assumption being true after observing the data – the posterior probability of manifest monotonicity –, this probability greatly depends on the prior probability that was assigned. In our approach, by using a uniform prior for each of the conditional probabilities, the probability that manifest monotonicity holds for a specific item decreases exponentially as the number of items included in the manifest score increases. A case could also be made for the proposition that since test items are artifacts constructed with the specific purpose of monotonically measuring a specific trait, manifest monotonicity is not extremely unlikely, even when a large number of items is included in the manifest score. However, the proposed procedure made use of the 'uninformative'

uniform prior distribution, giving each possible ordering of the conditional probabilities an equal probability. This was done with the idea in mind that the measure of support should solely reflect the extent to which the data (and not prior preconceptions of the researcher) point in the direction of manifest monotonicity. We contend that this is congruent with the idea that model assumptions should be critically evaluated, and that concerns raised about this assumption should be eliminated not by indicating that items were meant to be monotone by the person who designed them, but rather by having a careful look at the extent to which the data support this claim. This is precisely what the proposed procedure aims to do.

# Chapter 4

# Invariant Ordering of Item-Total Regressions

A new observable consequence of the property of invariant item ordering is presented, which holds under Mokken's double monotonicity model for dichotomous data. The observable consequence is an invariant ordering of the item-total regressions. Kendall's measure of concordance $W$ and a weighted version of this measure are proposed as measures for this property. Karabatsos and Sheu proposed a Bayesian procedure (2004), which can be used to determine whether the property of an invariant ordering of the item-total regressions should be rejected for a set of items. An example is presented to illustrate the application of the procedures to empirical data.

## 4.1 Introduction

Mokken's double monotonicity (DM) model (1971) is a nonparametric item response theory (IRT) model for the ordinal measurement of a single latent variable by means of a set of dichotomously scored items. The DM model

This chapter has been published as: Tijmstra, J., Hessen, D. J., Heijden, P. G. M., & Sijtsma, K. (2011). Invariant ordering of item-total regressions. *Psychometrika, 76*, 217–227.

is characterized by its definition of the item response function (IRF), which relates the probability of giving a positive or correct response to the latent variable. Unlike in parametric IRT models, such as the Rasch model (1960) and the Birnbaum models (1968), in the DM model the IRF is not parametrically defined. Instead, only order constraints are placed on the IRFs of a set of items. In contexts where the assumption of a parametric definition of the IRF is questionable or for purposes where an ordinal measurement level is sufficient, the DM model may be preferred over parametric IRT models.

Four assumptions define the DM model (Mokken, 1971; Sijtsma & Molenaar, 2002). The first assumption is unidimensionality, stating that the items measure one common latent variable. The second assumption restricts the item scores to be independent given the latent variable, and is known as the assumption of local independence (LI). The third assumption is latent monotonicity, stating that each IRF is a monotone nondecreasing function of the latent variable. The fourth assumption specifies that IRFs are nonintersecting, which is also known as invariant item ordering (IIO; Sijtsma & Junker, 1996).

The assumption of IIO distinguishes the DM model from the monotone homogeneity model (Mokken, 1971), in which IIO is not assumed, but which does assume unidimensionality, LI and latent monotonicity. Sijtsma and Junker (1996) discuss a variety of situations in which IIO is desirable or even necessary, such as the use of starting and stopping rules in intelligence testing based on the order of the item difficulties, or the analysis of differential item functioning. They point out that the property is also useful in the context of person fit analysis, where IIO can greatly facilitate the detection of aberrant response patterns.

More generally, the absence of IIO may complicate the interpretation of test results. For example, if a particular item is more difficult for high-ability subjects than another item, but easier for low-ability subjects, it may be difficult to provide an explanation as to why this difference in item ordering exists. For this reason, having IIO facilitates the interpretation of test results. Thus, IIO may be desirable when designing items for a test, and could also be used as a criterion for selecting items for a test during test construction.

Some procedures have been proposed that produce an overall measure of IIO. For example, Sijtsma and Meijer (1992) proposed to evaluate IIO using scalability coefficient $H^T$, which is based on Loevinger's $H$. The authors suggest guidelines to determine whether IIO should be rejected for a test based on this measure, but the level of significance and the power of this procedure are hard to establish. Ligtvoet, Van der Ark, Te Marvelde and Sijtsma (2010) further extended this procedure. Alternatively, Scheiblechner (2003) proposed to evaluate IIO using the weak order index $\sigma$, based on Goodman and Kruskal's $\gamma$ (Goodman & Kruskal, 1954). A drawback of this approach for testing for IIO is that it investigates evidence in favor of IIO (i.e., against independence of the item scores), and does not aim at testing for violations of IIO.

Rosenbaum (1987b) proved that, given LI, IIO of items $i$ and $j$ implies a manifest IIO of those items in any subpopulation that can be specified using the remaining items. Rosenbaum proposed to group subjects based on their unweighted sumscore on the remainder of the items. However, other groupings have also been proposed. Mokken (1971) proposed to use the score on a single item to group subjects. Another option is to use a vector containing a subset of the item scores on the remaining items, and test for IIO by investigating whether increasingness in transposition holds, which is implied by IIO (see Rosenbaum, 1987b; Sijtsma & Junker, 1996). If this property holds, observing a response vector with a score of 1 on an easy item and a score of 0 on a more difficult item should always be at least as probable as observing a response vector in which the more difficult item receives a 1 and the easier item a 0.

The drawback of methods that use subgroupings based on one or more of the remaining items is that the conclusion whether IIO holds for a set of items depends on many partial results. This renders such procedures laborious and, more importantly, makes it difficult to conclude whether IIO holds for the test as a whole. For example, if one tests for IIO per item pair by making use of the restscore for the item pair under consideration (i.e., the unweighted sumscore on the remainder of the items), one has to evaluate all item pairs separately in order to evaluate IIO for the whole test, because the partitioning of subjects based on their restscores is likely

to differ for different item pairs.

The use of the unweighted *total* score instead of the restscore would remedy this problem of having too many partial results, since a partitioning based on the total score would not vary over different item pairs. Thus, using the total score to test for IIO would facilitate the evaluation of the ordering of all items simultaneously instead of having to deal with separate item pairs, and would provide a more efficient method for IIO assessment. Using the total score to test for IIO would amount to determining whether the item probabilities conditional on the total score – the so-called item-total regressions – have the same ordering at every level of the total score. Such an ordering constitutes a manifest version of IIO, that is, a manifest invariant item ordering (MIIO) over the total score. Note that although an MIIO could be investigated over a variety of manifest scores, such as the mentioned restscore, in the remainder of the article the term MIIO will solely be used to refer to an MIIO over the total score.

An approach that focuses on the total score was proposed by Karabatsos and Sheu (2004), who suggested a Bayesian procedure that can be used to determine whether it is likely that an invariant ordering of the item-total regressions holds. The procedure makes use of a Gibbs sampler and results in a posterior-predictive $p$-value, which indicates whether we should reject or retain the assumption that the item probabilities are invariantly ordered over the total score. However, although they present this procedure as providing a test for Mokken's DM model, they do not provide a proof that MIIO holds under the DM Mokken model. While it has already been established that an invariant ordering of the item-*rest* regressions holds under the DM Mokken model (Sijtsma & Junker, 1996), this has not yet been proven for the item-*total* regressions.

The current article presents a proof showing that the DM model for dichotomous data implies an invariant ordering of the item-total regressions – an MIIO over the total score. Therefore, statistical tests for an invariant ordering of the item-total regressions can be used to test for the DM model. This proof provides the basis for the method to test the DM model that was proposed by Karabatsos and Sheu (2004). In addition to the proof, two measures of the property of an invariant ordering of the item-total regres-

sions are proposed, both of which make use of Kendall's (1939) measure of concordance $W$. The method of Karabatsos and Sheu is discussed as well, since with that method a decision can be made to reject or retain MIIO. The application of the measures based on Kendall's $W$ and the Karabatsos and Sheu method is illustrated using two sets of empirical data.

## 4.2 Theorem and Proof

Let $X_i$ denote the random variable for the score on item $i$, with realization $x_i$. Let $\mathbf{X} = (X_1, \ldots, X_k)$ denote the vector of item-score variables on a test with $k$ dichotomous items, with realization $\mathbf{x} = (x_1, \ldots, x_k)$. The latent variable is denoted by $\theta$, and the IRF of item $i$ is denoted by $P(X_i = 1|\theta) = P_i(\theta)$. One of the assumptions of the DM model is LI of the item scores given $\theta$, which can be presented as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P(X_i = x_i|\theta). \tag{4.1}$$

The assumption of IIO states that the items can be ordered such that

$$P_1(\theta) \leq \ldots \leq P_k(\theta), \qquad \text{for all } \theta, \tag{4.2}$$

where the indices $1, \ldots, k$ are assigned to the items based on decreasing difficulty. Furthermore, the total score $\sum_{i=1}^{k} X_i$ is denoted by $T$, and its realization by $t$. It is the claim of the present paper that (4.1) and (4.2) together imply the observable consequence of MIIO. This is stated in the following theorem.

*Theorem.* LI in (4.1) and IIO in (4.2) together imply

$$P(X_1 = 1\,|\,T = t) \leq \ldots \leq P(X_k = 1\,|\,T = t), \quad \text{for all } t, \tag{4.3}$$

where $P(X_i = 1\,|\,T = t)$ is the item-total regression of item $i$, that is, the probability of a positive response to item $i$ given $T = t$.

*Proof.* Following Hessen (2005), the probability of $\mathbf{X} = \mathbf{x}$ can be written as

$$P(\mathbf{X} = \mathbf{x} \,|\, \theta) = \left\{ \prod_{i=1}^{k} Q_i(\theta) \right\} \left\{ \prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i} \right\} V_k(\theta)^t,$$

where $Q_i(\theta) = 1 - P_i(\theta)$, $\omega_{ik}(\theta) = \frac{P_i(\theta)Q_k(\theta)}{Q_i(\theta)P_k(\theta)}$, and $V_k(\theta) = \frac{P_k(\theta)}{Q_k(\theta)}$. Any of the $k$ items can serve as the reference item. So, here, the choice of item $k$ as the reference item is arbitrary. Let $A_t$ be the set of all item-score vectors with total score $t$; that is, $A_t = \left\{ \mathbf{x} : \sum_{i=1}^{k} x_i = t \right\}$. Then, we may write

$$P(T = t \,|\, \theta) = \left\{ \prod_{i=1}^{k} Q_i(\theta) \right\} \left\{ \sum_{A_t} \prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i} \right\} V_k(\theta)^t,$$

and

$$P(\mathbf{X} = \mathbf{x} \,|\, T = t, \theta) = \frac{\prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}}{\sum_{A_t} \prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}}. \tag{4.4}$$

Also, let $B_{tx_jx_k}$ be the subset of $A_t$ in which the scores on items $j$ and $k$ are $x_j$ and $x_k$, respectively, and where $j$ is an arbitrarily selected item. That is, $B_{tx_jx_k} = \left\{ \mathbf{x} : x_j, x_k, \sum_{i=1}^{k} x_i = t \right\}$. Using (4.4), it follows that

$$P(X_j = x_j, X_k = x_k \,|\, T = t, \theta) = \frac{\omega_{jk}(\theta)^{x_j} \sum_{B_{tx_jx_k}} \prod_{i \neq j}^{k-1} \omega_{ik}(\theta)^{x_i}}{\sum_{A_t} \prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}},$$

for $i = 1, \ldots, k-1$, all $t$ and $\theta$. Note that $\sum_{B_{tx_jx_k}} \prod_{i \neq j}^{k-1} \omega_{ik}(\theta)^{x_i}$ is the same for $\left\{ \mathbf{x} : x_j = 1, x_k = 0, \sum_{i=1}^{k} x_i = t \right\}$ and $\left\{ \mathbf{x} : x_j = 0, x_k = 1, \sum_{i=1}^{k} x_i = t \right\}$. Hence,

$$\frac{P(X_j = 1, X_k = 0 \,|\, T = t, \theta)}{P(X_j = 0, X_k = 1 \,|\, T = t, \theta)} = \omega_{jk}(\theta),$$

for $j = 1, \ldots, k-1$, all $t$ and $\theta$. Since items $j$ and $k$ were selected arbitrarily, it follows that

$$P(X_i = 1, X_j = 0 \,|\, T = t, \theta) = P(X_i = 0, X_j = 1 \,|\, T = t, \theta)\, \omega_{ij}(\theta),$$

for all $t$ and $\theta$, and for any two items $i$ and $j$. Now, IIO implies that for any item pair $i$ and $j$, either $\omega_{ij}(\theta) \geq 1$ or $\omega_{ij}(\theta) \leq 1$, for all $\theta$. So if we arbitrarily let $\omega_{ij}(\theta) \geq 1$ for all $\theta$, then

$$P(X_i = 1, X_j = 0 \,|\, T = t, \theta) \geq P(X_i = 0, X_j = 1 \,|\, T = t, \theta), \text{for all } t \text{ and } \theta.$$

Adding $P(X_i = 1, X_j = 1 \,|\, t, \theta)$ to both sides gives

$$P(X_i = 1 \,|\, T = t, \theta) \geq P(X_j = 1 \,|\, T = t, \theta),$$

for all $t$ and $\theta$. Averaging both sides with respect to an arbitrary conditional density function $f(\theta \,|\, T = t)$ yields

$$\int P(X_i = 1 \,|\, T = t, \theta) f(\theta \,|\, T = t)\, d\theta \geq \int P(X_j = 1 \,|\, T = t, \theta) f(\theta \,|\, T = t)\, d\theta, \text{for all } t,$$

resulting in

$$P(X_i = 1 \,|\, T = t) \geq P(X_j = 1 \,|\, T = t), \quad \text{for all } t.$$

Hence, if IIO holds, then Equation 4.3 holds. This completes the proof.

## 4.3 Evaluating an Invariant Ordering of the Item-Total Regressions

Equation 4.3 specifies an order restriction on the item probabilities for each value of the total score. If for each value of the total score we investigate the ordering of the proportions of positive responses that were observed in the data, information can be obtained as to whether or not MIIO is violated. Table 4.1 displays these proportions of positive responses, denoted $p_{it}$, which are obtained by dividing the number of positive responses on item $i$ at level $t$ (denoted $s_{it}$) by the total number of responses at level $t$ (denoted $n_t$).

Note that since for $t = 0$ and $t = k$ by definition all items have the same proportion of success, the ordering at those levels does not contain

information regarding MIIO and hence need not be considered. The extent to which the orderings of $p_{it}$ $(i = 1, \ldots, k)$ differ for different values of $t$ $(t = 1, \ldots, k-1)$ gives a rough picture of whether the data indicate that MIIO is violated. In the next subsection, it is shown how this information can be summarized using two simple measures that are based on Kendall's (1939) measure of concordance $W$. Since the theoretical null distribution of these measures under MIIO is unknown, actually testing for MIIO using these measures is not an option that is available. In the second subsection, the statistical testing procedure proposed by Karabatsos and Sheu is discussed. This procedure results in a decision to retain or reject MIIO, and hence, whether the data contain evidence that IIO should be rejected.

Table 4.1: Observed proportions of positive responses

| | $t$ | | |
|---|---|---|---|
| Item | 1 | $\cdots$ | $k-1$ |
| 1 | $p_{11}$ | $\cdots$ | $p_{1(k-1)}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $k$ | $p_{k1}$ | $\cdots$ | $p_{k(k-1)}$ |

### 4.3.1 Kendall's $W$

A measure of the correspondence between the orderings of the item probabilities at different levels of the total score provides insight into whether MIIO is violated. For this purpose, we propose the use of Kendall's measure of concordance (Kendall's $W$; Kendall & Babington Smith, 1939). This nonparametric measure is related to the Spearman rank correlation coefficient (Spearman, 1904), but unlike the latter, $W$ can compare more than two orderings simultaneously. It can therefore be used to evaluate the degree of correspondence or concordance between the orderings of the items at different levels of the total score.

Kendall's $W$ takes on values between 0 (no correspondence) to 1 (perfect ordinal correspondence), and is usually employed to compare the orderings of the ratings of different judges. Kendall's $W$ is calculated on the basis of the rankings provided by a number of raters who independently ordered a number of objects. In the context of evaluating MIIO, these raters are replaced by the different levels of the total score (excluding $t = 0$ and $t = k$), and the objects by the items. For example, if the fictional proportions displayed in Table 4.2 are taken as the result of the ratings, they can be ordered for each level of the total score, which results in the rankings in Table 4.3.

Table 4.2: Proportions of positive responses at different levels of the total score, based on fictional data.

| | $t$ | | | |
|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 |
| 1 | .10 | .19 | .38 | .63 |
| 2 | .08 | .44 | .46 | .69 |
| 3 | .20 | .21 | .58 | .81 |
| 4 | .14 | .31 | .65 | .88 |
| 5 | .48 | .85 | .92 | 1.00 |

To determine the value of Kendall's $W$, the sum of the rankings of each item needs to be determined. Let $R_{it}$ be the rank of item $i$ obtained for total score $t$, where this ranking is based on increasing proportions. The sum of these rankings for item $i$ is obtained through

$$SR_i = \sum_{t=1}^{k-1} R_{it}. \tag{4.5}$$

Let $\mathbf{SR} = (SR_1, \ldots, SR_k)$. We consider the deviations of the elements of $\mathbf{SR}$ from their average value. Let $S$ represent the sum of squares of these

Table 4.3: Order of the proportions of positive responses at different levels of the total score, corresponding to Table 4.2.

| | $t$ | | | | |
|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | $SR_i$ |
| 1 | 2 | 1 | 1 | 1 | 5 |
| 2 | 1 | 4 | 2 | 2 | 9 |
| 3 | 4 | 2 | 3 | 3 | 12 |
| 4 | 3 | 3 | 4 | 4 | 14 |
| 5 | 5 | 5 | 5 | 5 | 20 |

deviations, then $W$ can be calculated using

$$W = \frac{12S}{(k-1)^2(k^3-k)}. \tag{4.6}$$

For the example in Table 4.3 one obtains

$$W = \frac{12 \times 126}{16(125-5)} = .7875.$$

One possible drawback of using Kendall's $W$ to measure the extent to which the data are in accordance with MIIO is that it does not take the number of observations into account that are available at the different levels of the total score. Kendall's $W$ expresses the degree of correspondence between the $k-1$ item orderings, and each level receives equal weight. However, when relatively few observations are available at a specific level of the total score, observing a discordant item ordering due to chance is more likely.

The different levels of the total score can be weighted by taking into account the number of persons that have the same total score $t$, denoted

by $n_t$. This can be accomplished by reformulating Equation 4.5 as

$$SR_i = \frac{k-1}{\sum_{t=1}^{k-1} n_t} \sum_{t=1}^{k-1} n_t R_{it}.$$

That is, the rankings that constitute $SR_i$ are weighted by $n_t$, and the value of $W$ under this weighting approach can be calculated as described in Equation 4.6. This weighted version of Kendall's $W$ is expected to be less sensitive to random fluctuations in the orderings.

If the sample value of $W$ or its weighted counterpart equals 1, there is no evidence available that MIIO is violated. If the value is smaller than 1, one can conclude that the ordering of the proportion of successes is not the same at every level of the total score. However, since deviations from perfect correspondence could be due to chance, one cannot determine whether MIIO is violated based on $W$ alone. To determine whether MIIO should be rejected, information is needed about the likelihood of obtaining a value as extreme (i.e., as low) as the observed value, given that MIIO holds.

Unfortunately, the only null distribution of $W$ or the weighted version of $W$ that is available is the distribution that corresponds to independence of the item orderings at the different levels of the total score. This null distribution enables one to test whether the item orderings show more invariance than one would expect to find under independence. However, if one wants to test for *violations* of MIIO, another null distribution is needed, one that corresponds to the situation where the item orderings are invariant over the total score. Regrettably, $W$ does not have a theoretical null distribution corresponding to the assumption of MIIO, and hence there is no exact test available that can be used to determine the probability of obtaining a value as low as the observed value. Thus, the two measures provide a useful way of summarizing the extent to which the data are in accordance with MIIO, but they do not provide an exact test to decide whether the property of MIIO – and hence the DM model – should be rejected.

### 4.3.2 Karabatsos and Sheu's Posterior-Predictive $p$-Value

To test for MIIO, Karabatsos and Sheu (2004) proposed using the table with observed proportions (Table 4.1), and determining how likely these proportions are when one assumes that MIIO holds. Let $\mathbf{p}$ denote the $k \times (k-1)$ matrix containing the observed proportions of positive responses as in Table 4.1; that is

$$\mathbf{p} = (p_{it} \,|i = 1, \ldots, k; t = 1, \ldots, k-1\,),$$

where $p_{it}$ denotes the observed proportion corresponding to item $i$ and a total score $t$. Let $\mathbf{\Delta}$ denote the $k \times (k-1)$ matrix containing the item-total regressions; that is

$$\mathbf{\Delta} = (P_{it}|i = 1, \ldots, k; t = 1, \ldots, k-1),$$

where $P_{it} = P(X_i = 1|T = t)$, and hence $\mathbf{\Delta} \in (0,1)^{k(k-1)}$. The likelihood of the data $\mathbf{p}$ given $\mathbf{\Delta}$ can then be assumed to be a product of $k(k-1)$ independent binomial probability mass functions:

$$L(\mathbf{p}|\mathbf{\Delta}) = \prod_{i=1}^{k} \prod_{t=1}^{k-1} \binom{n_t}{s_{it}} P_{it}^{s_{it}} (1 - P_{it})^{n_t - s_{it}}. \tag{4.7}$$

Let $\pi(\mathbf{\Delta})$ denote the prior distribution of the item probabilities in $\mathbf{\Delta}$. Let $\mathbf{\Omega}$ denote the subset of $(0,1)^{k(k-1)}$ that is in accordance with Equation 4.3; that is, the set of all matrices $\mathbf{\Delta}$ for which MIIO holds. The order constraints that follow from MIIO restrict this prior distribution in the following way,

$$\pi(\mathbf{\Delta}) \begin{cases} > 0 & \text{iff } \mathbf{\Delta} \in \mathbf{\Omega} \\ = 0 & \text{iff } \mathbf{\Delta} \notin \mathbf{\Omega} \end{cases}. \tag{4.8}$$

By combining Equations 4.7 and 4.8, the order-constrained posterior distribution of $\mathbf{\Delta}$ can be obtained through

$$\pi(\mathbf{\Delta}|\mathbf{p}) = \frac{L(\mathbf{p}|\mathbf{\Delta})\pi(\mathbf{\Delta})}{\int_{\mathbf{\Omega}} L(\mathbf{p}|\mathbf{\Delta})\pi(\mathbf{\Delta}) d\mathbf{\Delta}}. \tag{4.9}$$

Equation 4.9 cannot be evaluated analytically, but one can make use of the unconstrained posterior distribution of $\boldsymbol{\Delta}$,

$$\pi(\boldsymbol{\Delta}_u|\mathbf{p}) = \frac{L(\mathbf{p}|\boldsymbol{\Delta}_u)\pi(\boldsymbol{\Delta}_u)}{\int L(\mathbf{p}|\boldsymbol{\Delta}_u)\pi(\boldsymbol{\Delta}_u)d\boldsymbol{\Delta}_u},$$

where the subscript $u$ indicates that $\boldsymbol{\Delta}$ is no longer constrained to be part of $\boldsymbol{\Omega}$. This unconstrained posterior distribution can be modeled in terms of a beta distribution. Here, one has to use a beta density prior $\pi(\boldsymbol{\Delta}_u)$ in which each probability $P_{it}$ is specified independently and without the restriction of MIIO.

Using a Gibbs sampler (for details, see Karabatsos & Sheu, 2004), a large number of samples $(\boldsymbol{\Delta}^r|r = 1, \ldots, R)$ can be generated from the order-constrained posterior distribution $\pi(\boldsymbol{\Delta}|\mathbf{p})$. After discarding a proper burn-in period, $1, \ldots, B$, these draws results in an approximation of the posterior distribution $\pi(\boldsymbol{\Delta}|\mathbf{p})$.

To test whether MIIO is violated, the observed proportions $\mathbf{p}$ can be compared to the posterior-predictive distribution,

$$\pi(\mathbf{p}^{rep}\,|\mathbf{p}\,) = \int_{\boldsymbol{\Omega}} \pi(\mathbf{p}^{rep}\,|\boldsymbol{\Delta}\,)\pi(\boldsymbol{\Delta}\,|\mathbf{p}\,)d\boldsymbol{\Delta}.$$

This latter distribution can be approximated using the same Gibbs sampler, since after each iteration $r$, $\boldsymbol{\Delta}^{(r)}$ can be used to generate a new set of data, $\mathbf{p}^{rep}$. The posterior-predictive distribution is then approximated by the set of $\mathbf{p}^{rep}$ obtained in the Gibbs sampler.

To use this posterior-predictive distribution to test for violations of MIIO, Karabatsos and Sheu (2004) proposed using a chi-square discrepancy measure, defined as

$$\chi^2(\mathbf{p};\boldsymbol{\Delta}) = \sum_{i=1}^{k}\sum_{t=1}^{k-1}\left[\frac{(n_t p_{it} - n_t P_{it})^2}{n_t P_{it}}\right].$$

Using this measure, it is possible to obtain the posterior-predictive $p$-value (ppp-value):

$$\begin{aligned} p(\mathbf{p}|\boldsymbol{\Delta}) &= Pr\left[\chi^2(\mathbf{p}^{rep};\boldsymbol{\Delta}) \geq \chi^2(\mathbf{p};\boldsymbol{\Delta})|\mathbf{p}\right] \\ &= \int\int_{\boldsymbol{\Omega}} I\left[\chi^2(\mathbf{p}^{rep};\boldsymbol{\Delta}) \geq \chi^2(\mathbf{p}|\boldsymbol{\Delta})\right] p(\mathbf{p}^{rep}|\boldsymbol{\Delta})p(\boldsymbol{\Delta}|\mathbf{p})d\mathbf{p}^{rep}d\boldsymbol{\Delta}, \end{aligned}$$

where $I$ is an indicator function that equals 1 when the data $\mathbf{p}$ show less discrepancy relative to $\boldsymbol{\Delta}$ than $\mathbf{p}^{rep}$. This ppp-value indicates how likely it is to observe data as extreme as $\mathbf{p}$, under the assumption that MIIO holds. It can approximated using the samples $(\boldsymbol{\Delta}^r | r = B+1, \ldots, R)$, resulting in

$$\frac{1}{R-B} \sum_{r=B+1}^{R} I \left\{ \chi^2 \left( \mathbf{p}; \boldsymbol{\Delta}^{(r)} \right) \geq \chi^2 \left( \mathbf{p}^{rep}; \boldsymbol{\Delta}^{(r)} \right) \right\}.$$

This way, by selecting large enough values for both $B$ and $R$, an approximation of the ppp-value can be obtained, which indicates whether MIIO should be rejected. In order to decide whether MIIO should be rejected, a critical value needs to be selected for this ppp-value, for which Karabatsos and Sheu suggest the value of .15. Thus, by applying the Gibbs sampler it is possible to test for MIIO using a Bayesian approach.

## 4.4 Application to Empirical Data

The two measures based on Kendall's $W$ and the Karabatsos and Sheu procedure were used to evaluate MIIO for two sets of empirical data from a study in developmental psychology. One scale measuring nonaggressive behavior (7 items) and another scale measuring aggressive antisocial behavior (5 items) in male adolescents (Dekovic, 2003) were analyzed. Both scales consisted of polytomous items, but a dichotomization was easy to obtain, since each item measured the occurrence of specific types of antisocial behavior during the past year: to dichotomize the items, subjects received a score of 1 if the behavior had occurred, and a score of 0 otherwise. The sample size was 504, but due to missing values 8 subjects were excluded from the analysis of the nonaggressive antisocial behavior scale, and 6 subjects were excluded from the analysis of the aggressive antisocial behavior scale.

Table 4.4 and Table 4.5 provide a description of the items, with the corresponding overall sample proportions of positive responses and the sample proportions of positive responses for each level of the total score. The tables show that scores of 1 were obtained more often on the items of the

nonaggressive antisocial behavior scale than on the items of the aggressive antisocial behavior scale, and that the overall proportion of positive responses varied more between the items on the former scale than on the latter one, indicating a larger spread in item difficulties on the nonaggressive behavior scale.

Table 4.4: Overall Proportions of positive responses and proportions conditional on $t$ for the nonaggressive antisocial behavior scale.

| | Proportion | $t$ | | | | | |
|---|---|---|---|---|---|---|---|
| Item Description | Positive | 1 | 2 | 3 | 4 | 5 | 6 |
| 1: Disregarding parents | .69 | 0.59 | 0.78 | 0.92 | 0.91 | 0.96 | 1.00 |
| 2: Missing a curfew | .44 | 0.15 | 0.34 | 0.67 | 0.84 | 0.89 | 1.00 |
| 3: Skipping school | .16 | 0.02 | 0.09 | 0.18 | 0.25 | 0.50 | 0.59 |
| 4: Cheating on a test | .43 | 0.14 | 0.42 | 0.55 | 0.79 | 0.96 | 0.93 |
| 5: Fare dodging | .26 | 0.05 | 0.18 | 0.27 | 0.58 | 0.57 | 0.93 |
| 6: Shoplifting | .24 | 0.03 | 0.15 | 0.25 | 0.44 | 0.75 | 0.86 |
| 7: Stealing from someone | .14 | 0.03 | 0.04 | 0.16 | 0.19 | 0.36 | 0.69 |

Table 4.5: Overall proportions of positive responses and proportions conditional on $t$ for the aggressive antisocial behavior scale.

| | Proportion | $t$ | | | |
|---|---|---|---|---|---|
| Item Description | Positive | 1 | 2 | 3 | 4 |
| 1: Fire setting | .10 | 0.10 | 0.19 | 0.38 | 0.63 |
| 2: Carrying a weapon | .27 | 0.48 | 0.85 | 0.92 | 1.00 |
| 3: Threatening with a weapon | .12 | 0.08 | 0.44 | 0.46 | 0.69 |
| 4: Beating someone | .14 | 0.14 | 0.31 | 0.65 | 0.88 |
| 5: Street fighting | .13 | 0.20 | 0.21 | 0.58 | 0.81 |

For both scales, the two measures based on $W$ were calculated, and the

Bayesian procedure was used to test whether MIIO should be rejected. For the nonaggressive antisocial behavior scale, we found $W = .923$, and the weighted version of $W$ equalled .938. These values are close to 1, suggesting little evidence that MIIO is violated. The ppp-value of .537 (based on 5,000 iterations) supports this conclusion, and hence MIIO was not rejected for this scale.

For the scale measuring aggressive antisocial behavior, $W = .788$, and the weighted $W$ equalled .731. These values are lower than the values obtained for the nonaggressive scale. The Bayesian procedure resulted in a ppp-value of .142, which is just below the critical value of .15. Thus, the results suggest that for this scale MIIO appears to be violated. Hence, IIO and the DM model can be rejected for this scale.

Thus, for the nonaggressive antisocial behavior scale, MIIO was not rejected, and hence IIO need not be rejected for this scale. If IIO does indeed hold for this scale, this means that the different types of behavior measured by the nonaggressive scale come in a specific order, with one kind of behavior always being more likely than another, regardless of how antisocial the adolescent is. This could be an interesting substantive finding, resulting from the non-rejection of IIO. Likewise, it would be interesting to know why IIO does not appear to hold for the aggressive antisocial types of behavior. Perhaps some types of behavior display nonmonotonicities, only being exhibited frequently by mildly antisocial youths. Again, such a conclusion based on IIO research could result in relevant substantive considerations.

## 4.5 Conclusion and Discussion

For a test consisting of dichotomous items, it was shown that under LI the property of IIO over the latent variable implies the property of manifest invariant ordering of the item-total regressions; that is, MIIO over the total score. This result implies that MIIO not only holds for the Rasch model (Hessen, 2005), but also for the DM model (Mokken, 1971). Thus, investigating MIIO is not only useful in the context of parametric IRT, but

also in the context of nonparametric IRT. Inspection of MIIO is relatively simple, and may be an attractive method in IIO research. This way, the theorem provided in this article helps to facilitate IIO research.

The two measures of MIIO, both based on Kendall's (1939) measure of concordance $W$, reflect the extent to which the data are in accordance with MIIO. Values of 1 imply that there is no evidence available that MIIO is violated. Values lower than 1 show that the data are not completely in accordance with MIIO. Karabatsos and Sheu (2004) proposed a Bayesian procedure to determine whether it is likely that MIIO is violated. This Bayesian procedure results in a decision to reject or retain MIIO for a test consisting of dichotomous items. In addition to this Bayesian approach, it would be interesting to investigate whether it is also possible to provide a frequentist test, perhaps making use of the constrained statistical inference framework (see, e.g., Silvapulle & Sen, 2005).

Regardless of which testing procedure is used, a rejection of MIIO implies a rejection of IIO, and hence testing procedures for MIIO can be used to determine whether the application of IRT models that assume IIO, such as the DM model and by implication the Rasch model, would be appropriate. Furthermore, IIO is an attractive property in itself that one could pursue during test construction, since it allows for an unambiguous ordering of the items based on their difficulty, which makes interpretation of the test results easier. Additionally, IIO may be required in applications where starting and stopping rules need to be applied, or where the items need to be presented in order of difficulty (Sijtsma & Junker, 1996). Differential item functioning analysis and person fit analysis may also benefit from having IIO.

Knowing whether IIO holds can also be of substantive importance. Whenever IIO is violated for two specific items, groupings of respondents can be made for which the order of the probabilities of these items is reversed. This will at the very least require an explanation, telling us why it is the case that, for example, a certain item is relatively easy for high ability examinees, perhaps pointing to some new insight that helps them deal with an item that would otherwise be relatively difficult. As shown in the example, it might also point to the possibility that some behavior

ceases to become more frequent after a certain level of the latent variable has been reached, perhaps even indicating nonmonotonicities. For these reasons, determining whether MIIO and hence IIO should be rejected can be of considerable importance.

# Chapter 5

# Why We Need to Assess Prior Plausibility when Evaluating Statistical Model Assumptions

This article explores whether the null hypothesis statistical testing (NHST) framework provides a sufficient basis for the evaluation of statistical model assumptions. It is argued that while NHST-based tests can provide some degree of confirmation for the model assumption that is evaluated – formulated as the null hypothesis –, these tests do not inform us of the degree of support that the data provide for the null hypothesis and to what extent the null hypothesis should be considered to be plausible after having taken the data into account. Addressing the prior plausibility of the model assumption is unavoidable if the goal is to determine how plausible it is that the model assumption holds. Without assessing the prior plausibility of the model assumptions, it remains uncertain whether the model of interest gives an adequate description of the data and thus whether it can be considered valid for the application at hand. Although addressing the prior plausibility is difficult, ignoring the prior plausibility is not an option

if we want inferences based on the model to have some degree of plausibility. A case study from item response theory is presented to illustrate the importance of taking prior plausibility into account.

## 5.1 Introduction

Whenever statistical models are used to make inferences, the question whether these inferences are justified is important. Whether we are trying to predict someone's future income using information about their educational level or whether we want to estimate someone's intelligence based on a set of responses, the correctness of our statistical inferences depends on whether we have correctly specified the statistical model. For a conclusion based on a statistical model to be valid, all the assumptions needed to reach that conclusion have to hold. If one of these assumptions is violated the model is formally invalid and inferences based on that model lose credibility, with more credibility being lost if the model is not robust against such a violation.

Since every statistical model has to rely on some assumptions for its inferences, the evaluation of these statistical model assumptions is an inevitable and crucial part of statistical modeling. Some model assumptions can be made plausible through non-statistical means, but in most cases the evaluation of model assumptions requires the use of statistical methods.

Model assumptions can be treated as statistical hypotheses that can be evaluated using statistical testing procedures, sometimes supplemented by graphical methods (see e.g. Tabachnick & Fidell, 2001; Gibbons, 1985; Montgomery, Peck & Vining, 2001; Michael, 1983). The standard way of dealing with these statistical hypotheses is based on frequentist approaches, which mainly use the null hypothesis statistical testing (NHST) framework (see e.g. Fisher, 1956; Tabachnick & Fidell, 2001). Through the use of statistical tests, these procedures aim to give us sufficient information about the model assumption to be able to determine whether the assumption – formulated in the null hypothesis – can be maintained or whether it should be rejected.

This paper explores whether NHST-based approaches succeed in providing sufficient information to determine how plausible it is that a model assumption holds, and whether a model assumption is plausible enough to be accepted. The paper starts out with a short motivating example, which is aimed at clarifying the issues that arise when model assumptions are evaluated using null hypothesis tests.

In section 2, the background of the NHST framework and its methodology are discussed, in particular the way in which the framework is applied to evaluate model assumptions. It is argued that while the NHST framework normally focusses on falsifying the null hypothesis, the evaluation of model assumptions actually requires the framework to focus on verifying or confirming the null hypothesis. This results in a mismatch between the justification behind NHST and the way in which NHST is applied to evaluate statistical model assumptions.

Section 3 explores the extent to which a null hypothesis test can provide confirmation for the model assumption it evaluates. It is argued that while NHST-based procedures are able to show whether the data provide support in favor or against the model assumption, they do not inform us of the *strength* of this support. Furthermore, null hypothesis tests cannot inform us to what extent we should consider the model assumption to be *plausible* after having taken the data into consideration. It will be shown that this 'posterior plausibility' can only be evaluated if we take the plausibility of the model assumption before observing the data into consideration – the 'prior plausibility'. Because after applying null hypothesis tests to the data it still remains unclear how plausible the model assumptions are, it is concluded that the NHST framework is unable to inform us to what extent inferences made using the model can be trusted.

In section 4, the importance of taking the prior plausibility of model assumptions into consideration is illustrated using an example from item response theory, which is the evaluation of the assumption of latent monotonicity. Both the application of null hypothesis tests and that of a Bayesian alternative are discussed, as well as the issues that one may run into when assessing the prior plausibility of the assumption of latent monotonicity.

### 5.1.1 A Motivating Example

The issues that arise when model assumptions are evaluated using NHST can be illustrated using a hypothetical example. Consider a researcher who wants to find out whether it makes sense to develop different types of educational material for boys and for girls. For this purpose, he is interested in possible gender differences in the cognitive processing speed of children in the context of spatial reasoning. He has constructed a set of ten items that are assumed to measure this ability, and for each respondent records the total response time. He uses the total response time as a measure with which he investigates possible differences in processing speed between boys and girls. We will ignore all the issues that may arise concerning the validity of this measure and focus solely on the assumptions that are needed for his analysis.

Let us assume that the researcher uses Student's independent samples $t$-test to compare the average response speed of boys and girls. Before he can interpret the results of this $t$-test, he has to make sure that all the assumptions required by this test are met. In this case, he has to check whether both the response times of the boys and the response times of the girls are independently and identically distributed (i.i.d.), whether these two distributions both correspond to a normal distribution, and whether the variances of these distributions are the same (Tabachnick & Fidell, 2001; Field, 2009).

For reasons of simplicity, let us only consider the assumption of equal variances, assuming that the researcher has somehow established that the other two assumptions hold. The assumption of equal variances can be assessed roughly using graphical methods (i.e., by considering a plot that both distributions). Since it is difficult to form a precise conclusion about the assumption using graphical methods alone, the assumption is usually also tested using NHST-based methods, such as Levene's test (Levene, 1960). In this example, Levene's test can be used to evaluate the null hypothesis $H_0 : \sigma^2_{boys} = \sigma^2_{boys}$. Like all NHST procedures, Levene's test produces a $p$-value. This $p$-value tells the researcher how probable it is to obtain a difference in variances of the same size as or greater than the difference he

obtained in his samples, if the variances of the populations are the same. If the $p$-value is smaller than the prespecified level of significance $\alpha$ (and for Levene's test $\alpha$ is usually set to .05 or .01; Field, 2009), the null hypothesis is rejected, and the conclusion is drawn that the model assumption is not warranted.

Let us assume that in this case, Levene's test results in a $p$-value of .23, which is well above the level of significance. The researcher concludes that there is no reason to worry about the assumption of equal variances, and proceeds with applying the $t$-test. The results of the $t$-test tell him that boys are significantly faster in solving the spatial reasoning items ($p = .032$, $\alpha = .05$) (inference 1). When he considers the 95% confidence interval, he concludes that the observed difference should also be considered relevant (inference 2). Based on these two statistical inferences the researcher concludes that it may be useful to develop additional educational material to help the girls with spatial reasoning.

The present example is meant to exemplify how statistical model assumptions are commonly dealt with in practice (see also Gigerenzer, 2004). At the surface, the steps that the researcher takes in this example seem unproblematic. However, important questions can be asked about the justification of the inferences made in this example, and in cases like these in general. Should the researcher be worried about his model assumption if he has background knowledge that boys commonly show a larger variance in cognitive processing speed than girls? Can he safely conclude that the variances are indeed equal in the population and that the model assumption holds, or that this is at least very likely? And if there remains uncertainty about the plausibility of this model assumption, how should this uncertainty influence his statistical inferences?

Questions like these are not restricted to the application of $t$-tests, but apply equally strong to all areas where statistical models are used to make inferences. Since statistical inferences based on these models often form the basis for action, determining whether the inferences that are made using NHST can be trusted is of both theoretical and practical importance.

## 5.2 The Null Hypothesis Statistical Testing Framework

### 5.2.1 Background of the NHST Framework

The basis of the NHST framework goes back to the frequentist paradigm founded by Fisher in the 1930s (Fisher, 1930; 1955; 1956; 1960; Hacking, 1976; Gigerenzer, 1993), as well as the frequentist paradigm founded by Neyman and Pearson in that same period (Neyman, 1937; 1957; Pearson, 1955; Neyman & Pearson, 1967; Hacking, 1976; Gigerenzer, 1993). Whereas Neyman and Pearson proposed to contrast two hypotheses that are in principle on equal footing, Fisher's approach focusses on evaluating the fit of a single hypothesis to the data, and has a strong focus on falsification. Starting from the 1950s, elements from both approaches were incorporated in the hybrid NHST framework as it exists today in the social and behavioral sciences (Gigerenzer & Murray, 1987; Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989; Gigerenzer, 1993; Lehmann, 2006): While this framework proposes to evaluate a null hypothesis in contrast with an alternative hypothesis – in line with Neyman and Pearson –, the focus lies on attempting to reject the null hypothesis – in line with Fisher's methodology. Thus, despite the important differences that existed between the paradigms of Fisher and that of Neyman and Pearson (Gigerenzer, 1993), the current framework constitutes a hybrid form of the two paradigms (Gigerenzer et al., 1989; Gigerenzer, 1993).

The focus on trying to reject the null hypothesis in the current NHST framework is strongly tied to Popper's advocation of falsification as the only appropriate method of inference in the empirical sciences (Nickerson, 2000). Popper (1959) claimed that every scientific theory specifies certain events that cannot occur under that theory. By testing whether these events do occur, the possibility arises to conclusively falsify the theory, thus leading to scientific progress by dismissing a theory as being false.

While it is almost always possible to search for (and usually find) support for a scientific theory (Popper, 1974), the problem of induction (Hume, 1888) implies that no theory that makes universal statements about an infi-

nite set of objects can ever be fully confirmed through concrete observations. Most scientific theories deal with these kind of universal statements, and no finite number of observations will ever result in a verification of those theories, even if all observations so far are consistent with those theories. Thus, Popper argues, only falsification can provide us with certainty. By exposing implications of a scientific theory to critical tests, scientists should aim at falsifying their theories and thus improve our understanding of the universe by eliminating false theories.

If a theory has survived a critical test without being falsified, this theory is not confirmed or verified – complete verification being logically impossible –, but rather can be said to have been 'corroborated'. According to the strict view that Popper proposes, corroboration does not imply that the theory under consideration is now more likely to be true, since there are still an infinite number of instances that can result in the falsification of the theory. Thus, Popper claims that only falsification provides epistemologically relevant new information about the theory, since we now know for sure that it is false.

In statistical inference, the absolute implications of falsification are abandoned, since statistical conclusions are probabilistic rather than deductive. However, the focus on falsification is maintained in the NHST framework (Fisher, 1955; 1960; Mayo, 1996; Gigerenzer, 1993). By formulating a statistical null hypothesis and deriving the distribution of a test statistic under the null hypothesis, a statement can be made about the $p$-value: the probability of observing an outcome under the null hypothesis that is at least as extreme as the outcome that was observed. If the $p$-value is sufficiently low, it is concluded that the data are unlikely to be observed under the null hypothesis. If the $p$-value falls below a preset level of significance $\alpha$ the null hypothesis is rejected or falsified, and the alternative hypothesis is accepted (Neyman & Pearson, 1967). If $p \geq \alpha$, the null hypothesis is retained, but – consistent with Poppers methodology – no conclusions are drawn about the truth of the null hypothesis (Fisher, 1955). Mirroring Popper, in this case the testing procedure tells us that the null hypothesis is retained rather than confirmed.

Because the NHST framework explicitly does not address the probabil-

ity that the null hypothesis is true (Edwards, Lindman & Savage, 1963), NHST remains silent about the effect that a failure to reject the null hypothesis should have on our assessment of the plausibility of the null hypothesis – the extent or degree to which we believe that it is credible or likely to be true. Similar to the way in which Popper suggests to evaluate theories, null hypotheses are simply considered hypotheses that have not yet successfully been rejected, but should not receive any credibility.

### 5.2.2 Using Null Hypothesis Tests to Evaluate Model Assumptions

For the critical evaluation of substantive theories, only focussing on falsifying hypotheses may be defensible from a purely scientific perspective, in line with the philosophy of Popper (see e.g. Mayo, 1996). The theory should be put to a critical test, and this is something that can be done using the NHST framework. Since model assumptions can also be formulated as hypotheses, it seems attractive to make use of the NHST framework to evaluate these assumptions.

By formulating the statistical model assumption as a null hypothesis, NHST-based approaches attempt to expose the model assumption to a critical test in an attempt to falsify it, to determine whether applying the model might be inappropriate. If the test statistic is significant and the null hypothesis is rejected, it is concluded that the model is formally invalid and that inferences based on the model are not (or are not automatically) valid.[1] If the null hypothesis is not rejected, the model assumption has survived a critical test, and can be maintained.

The difficulty lies in the fact that the NHST-framework normally informs us that a nonrejection of the null hypothesis does not confirm the null hypothesis; it is only *maintained* and not *accepted*. After all, a Type II error could have occurred, so we could make a mistake in accepting $H_0$ to be true. However, this absence of confirmation in NHST is problematic in the case of model assumptions, since now we want to know whether the

[1] The issue of robustness is dealt with at the end of section 5.3.3.

null hypothesis can be accepted rather than maintained. That is, if a non-rejection of $H_0$ does not provide at least some degree of confirmation for $H_0$, then we still have no information about whether we can use the model, since for the model to be correct all of its assumptions have to hold.

This strict application of the NHST framework thus only informs us whether the model should be considered to be inappropriate and not whether we can assume that it is correct. While this position may be theoretically defensible, its implications are severe and do not match scientific practice. That is, using only NHST no model assumption would ever receive any degree of confirmation from the data, and their plausibility would always remain completely uncertain.

If we never have any positive evidence that our statistical model is correct, then it remains unclear why we should put any trust in conclusions based on that model. In our example, we might conclude that there is a significant difference between the response speed of boys and girls, but we would have to add that we are completely uncertain about whether that inference can be trusted. If this is the case, then it is unclear why we should put more trust in our statistical models than in simple subjective guesswork. Since claiming that we never have any evidence in favor of the model assumption would make all statistical inference arbitrary, it is assumed in the remainder of the paper that the application of the NHST framework to model assumptions only makes sense if it allows for some form of confirmation of the null hypothesis.

One way to avoid the implication that we cannot put any faith in our statistical models is to change the implication of a nonrejection of $H_0$: Instead of just maintaining $H_0$, we can decide to accept $H_0$ when it is not rejected. This would solve the problem of never concluding that our model assumptions hold. This approach seems to be implicitly embraced in practice, where researchers check their assumptions and then decide to continue to apply and trust their model if these tests do no indicate that there is a problem with the assumptions. The example presented earlier was also in line with this approach. Because this approach allows for the acceptance of $H_0$ and thus is more liberal than the strict version of NHST that was discussed before, we will call it the 'liberal approach' to NHST. For con-

trast, the strict interpretation of NHST that solely focuses on falsification and does not allow for any degree of confirmation of $H_0$ will be called the 'strict approach'.

Because the liberal approach moves away from the falsification-based foundations of the strict NHST framework, it raises the question why a nonrejection of $H_0$ should provide us with sufficient reason to accept $H_0$. NHST was not designed to provide confirmation for the null hypothesis, so it is important to examine whether the conclusion to accept $H_0$ after a nonrejection can be justified. The following section illustrates that the NHST-framework alone cannot provide us with sufficient information to justify the decision to accept $H_0$ after a nonrejection, and hence that the liberal approach cannot inform us whether our model inferences can be trusted.

Another possible response to the problems of the strict approach to NHST is to abandon the idea that a dichotomous decision needs to be made about the plausibility of the model assumption based on the $p$-value that is obtained. Since we know that both Type I and Type II errors can occur, it could be argued that drawing definite conclusions about the truth of the model assumption will always be premature. Rather than accepting or rejecting $H_0$, we might decide to always maintain $H_0$ and continue to apply the model, while being aware that $H_0$ might be false. It could be argued that we should interpret the significance of the test statistic as a dichotomous measure that provides us with some degree of support in favor or against the model assumption, and that we should incorporate this information in determining the extent to which we trust inferences based on the model. Because this approach does not result in a dichotomous decision about the model assumption, we will call it the 'continuous approach' to contrast it with both the strict and the liberal approach.

The continuous approach is more in line with the position that is defended in this paper, but the next section will illustrate that it runs into two problems. First, the $p$-value alone only provides ordinal information about the plausibility of the model assumption: $H_0$ either becomes more plausible (when $p \geq \alpha$) or less plausible (when $p < \alpha$), but it is not clear *how much more or less plausible* it has become after having performed the

null hypothesis test. Second, even if we could determine the extent to which the test result makes $H_0$ more or less plausible than it was before performing the test, it still remains unclear *how plausible* $H_0$ is, since this plausibility could still range from highly plausible to highly implausible. If it is unclear how plausible the model assumption is, it also remains unclear to what extent we should trust our model inferences.

The next section illustrates that both the liberal and the continuous approach run into problems when trying to assess the plausibility of model assumptions. It is argued that these problems of NHST in confirming the model assumption can only be solved if one takes background information about the plausibility of the model assumption into account. That is, the evaluation of model assumptions should incorporate an assessment of their prior plausibility. Although addressing this prior plausibility is difficult, it has to be addressed if we want inferences based on the statistical model in question have some degree of credibility.

## 5.3 Confirmation of Model Assumptions Using NHST

The previous section argued that if NHST is to inform us whether a model assumption has some degree of plausibility, NHST has to allow for some form of confirmation of the null hypothesis. Regardless of whether we want to make a dichotomous decision about our model assumption – accept or reject it – or whether we want to assess the degree to which it can be considered plausible, any approach that tries to address the plausibility of the model assumption has to assess the evidence in favor or against that assumption. This raises two questions: What should we count as relevant evidence about the plausibility of the model assumption, and how can we relate this evidence to the plausibility of the model assumption.

Both the liberal and the continuous approach to NHST provide the same answers to these two questions. In response to the first question the NHST framework tells us that the evaluation of hypotheses – and hence also our model assumptions – should occur based solely on the evidence available in the data (Gigerenzer, 1993). In response to the question of relating

the evidence to the plausibility of the model assumption, NHST concludes that all the relevant information about the plausibility of $H_0$ is captured by the $p$-value (Trafimow, 2003; Gigerenzer, 1993). While the liberal and the continuous approach differ in the conclusions they draw based on this $p$-value, they agree that conclusions about the plausibility of $H_0$ should be drawn solely on the basis of the $p$-value.

This section shows that it would be irrational to base our evaluation of the plausibility of the model assumption on the $p$-value alone if we already have some prior ideas about the plausibility of that model assumption. If we try to base our conclusions about the model assumption on the $p$-value alone, this may be at the cost of probabilistic inconsistency with our prior beliefs. Since in evaluating model assumptions our goal is to assess if these assumptions are plausible, ignoring relevant information that we may have about the truth of the model assumption would not be rational. Thus, neither the liberal nor the continuous approach to NHST provide us with rational guidelines to form our beliefs about the plausibility of model assumptions.

Although NHST does not provide defensible guidelines for assessing the plausibility of model assumptions based on the data alone, one could hope that it succeeds in capturing the extent to which the data provide support for the plausibility of the model assumption. That is, NHST might not succeed in determining *how plausible* $H_0$ is, but perhaps it can inform us *how much more or less plausible* it has become after having observed the data. However, this section illustrates that this claim is also problematic, since the degree of support that the $p$-value provides in favor or against the model assumption also depends on our prior assessment of the plausibility of $H_0$ and all of its possible alternatives. Before these issues can be addressed however, it is important to formalize the notion of prior and posterior plausibility.

### 5.3.1 Prior and Posterior Plausibility of the Model Assumption

To determine if the NHST framework can inform us about the plausibility of $H_0$, we have to clearly define what we take plausibility to be. Let us formalize the notion of plausibility by requiring it to take on a value that ranges from 0 (completely implausible or certainly wrong) to 1 (completely plausible or certainly right). This value represents the degree of plausibility that is assigned to a proposition, for example the proposition '$H_0$ is true'.

Since model assumptions are arguably either true or false, if we had complete information (and were able to process it), there would be no uncertainty about the model assumptions and we would assign a value of either 0 or 1 to the plausibility of a model assumption being true. However, researchers are forced to assess the plausibility of the assumption using incomplete information, and their assessment of the plausibility depends on the limited information that they have and the way in which they evaluate this information. Thus, when we talk about the plausibility of a model assumption, it will always be conditional on the person that is doing the evaluating and the information that he has.

Let us denote the plausibility of a hypothesis $H_0$ as it is assessed by a rational and coherent person $j$ by $P_j(H_0)$. Such rational and coherent persons may not actually exist, but they can be considered to at least serve as idealizations for the way in which we should revise our beliefs in the face of new evidence. Thus, $P_j(H_0)$ represents the degree to which person $j$ believes in the proposition '$H_0$ is true', and it therefore tells us to what extent person $j$ thinks that it is probable that $H_0$ is true. This 'degree of belief' is in the Bayesian literature on statistics and epistemology often called the 'subjective probability' or 'personal probability' that a person assigns to the truth of a proposition (see Savage, 1972; Howson & Urbach, 1989; Earman, 1992; Suppes, 2007). As a way of quantifying this subjective probability or degree of belief, we could imagine asking this person how many cents he would be willing to bet on the claim that $H_0$ is true if he will receive 1 dollar in the case that $H_0$ is indeed true (Ramsey, 1926; Gillies, 2000).

Let us assume that person $j$ has obtained a data set $\boldsymbol{X}$ – the data to which he hopes to apply the model – and that he wants to determine how plausible it is that $H_0$ holds after having taken the data into consideration. Let us call his prior assessment $P_j(H_0)$ of the plausibility of $H_0$ before considering the data $\boldsymbol{X}$ the *prior plausibility*. Since person $j$ wants to determine whether he should trust inferences based on the model, he wants to make use of the information in the data to potentially improve his assessment of the plausibility of $H_0$. Thus, to make a better assessment of how plausible $H_0$ is he wants to update his prior belief based on the information in the data. Let us call this assessment of the plausibility that has been updated based on the data $\boldsymbol{X}$ person $j$'s posterior plausibility, which we denote by $P_j(H_0|\boldsymbol{X})$.

Thus, the prior and the posterior plausibility correspond to subjective probabilities, which are assigned by the person who evaluates the model assumption. These subjective probabilities can be considered to differ from 'objective' probabilities that can be assigned to events, such as coin flips (Gillies, 2000). Adherents of the NHST framework often argue that because subjective probabilities deal with propositions rather than with events, these subjective probabilities are not 'real' probabilities (Mayo, 1996), while others argue that the two notions of probability can coexist (Gillies, 2000). Whether or not these subjective probabilities should indeed be called 'probabilities' or rather should be called 'degrees of belief' is not relevant to the point of this paper, but for convenience and for consistency with the Bayesian literature both the term 'subjective probability' and 'plausibility' are used in the remainder of the paper to describe this degree of belief. Whenever the terms 'prior probability' or 'posterior probability' are used, they refer to subjective probabilities.

### 5.3.2 Relevance of Prior Knowledge about the Model Assumption

Since both the liberal and the continuous approach to NHST posit that our evaluation of the plausibility of $H_0$ should be based only on the data (Trafimow, 2003), they tell us that our prior beliefs about the possible truth

of $H_0$ should be completely overridden by the information about the model assumption that the null hypothesis test provides us with. The idea is that this way, the influence of possible subjective considerations is minimized (Mayo, 1996). Proponents of the NHST-framework cannot allow the prior assessment of the plausibility of $H_0$ to influence the conclusions that are drawn about the plausibility of $H_0$ without abandoning the idea that the $p$-value contains all the relevant information about the plausibility of $H_0$. Hence, it will be assumed that if person $j$ follows NHST-based guidelines in assessing the plausibility of $H_0$, $P_j(H_0|\boldsymbol{X})$ will not depend on $P_j(H_0)$, but solely depends on the $p$-value.

However, there are clear cases where our assessment of the plausibility of $H_0$ *should* depend on our prior knowledge if we are to be consistent. If for some reason we already know the truth or falsehood of $H_0$, then basing our assessment of the plausibility purely on the result of a null hypothesis test – with the possibility of a Type I and Type II error, respectively – can only make our assessment of the plausibility of $H_0$ worse, not better. When we know in advance that a model assumption is wrong, failing to reject it should not in any way influence our assessment of the assumption.

For example, assume that due to a methodological flaw in the design of an experiment we know for sure that the responses of the subjects were not obtained independently (perhaps we saw one subject copy the answers of another subject). If we then statistically test the assumption of independence with a null hypothesis test, a nonsignificant result should not give us any confidence in that assumption, since we know that a Type II error has occurred. Likewise, obtaining a significant result when testing an assumption that we know to be true (perhaps because we know under which model the data were generated) should not influence our confidence in that assumption. In these extreme cases, the available background information should completely overrule the conclusions drawn by the null hypothesis tests.

More generally, one could conclude that the less plausible a null hypothesis is on the basis of the background information, the more hesitant we should be to consider it to be plausible if it fails to be rejected. The data may provide us with relevant information about the model assump-

tion, but this should influence our assessment of the plausibility of $H_0$ in a way that is consistent (i.e., adhering to the laws of probability) with our prior assessment of the plausibility of $H_0$.

Thus, our prior beliefs about the plausibility of the model assumption are relevant for our assessment of the plausibility of $H_0$ after having taken new data into account. While the prior plausibility $P_j(H_0)$ is a subjective probability, this does not mean that it is arbitrary or that it cannot be the result of an informed judgment or be subject to reasons (Howson & Urbach, 1989; Lee & Wagenmakers, 2005). For example, person $j$ might be worried about subjective biases, and based on the principle of indifference he may want to assign equal probability to $H_0$ and its alternative(s) (Jeffreys, 1961). Another possibility is that the person has good reason to doubt whether the model assumption is true. In the context of the motivating example, the researcher may be aware of previous research indicating that boys generally show greater variance in cognitive processing speed than girls. This could give him strong reasons to suspect that boys will also show greater variance on his particular measure than girls, and this background information would then be incorporated in his assessment of the plausibility of the assumption of equal variance before he considers the data.

Regardless of the specific value of $P_j(H_0)$ or the way in which person $j$ arrived at this particular value, he may want to revise his belief in the model assumption after having taken the data into account. Because he is rational, he will want to update his beliefs according to the laws of probability after having observed the data to come to a rational assessment of the plausibility of $H_0$. The next subsection illustrates that this assessment of the posterior plausibility of $H_0$ can only be consistent if it incorporates the prior assessment of the plausibility. That is, $P_j(H_0|\mathbf{X})$ should depend on $P_j(H_0)$.

### 5.3.3 Assessing the Plausibility of $H_0$ Using a Null Hypothesis Test

To examine how NHST may help to evaluate the plausibility of a statistical model assumption, let us further examine the hypothetical case of

researcher $j$ who wants to apply a statistical model to a data set $\boldsymbol{X}$, and who wants to evaluate one of the assumptions defining that model. For convenience, let us assume that the other assumptions of the model have already been established to hold. Let us also assume that the model assumption that he evaluates can be formulated as a simple null hypothesis, specifying that a parameter has a specific value. For example, this null hypothesis could correspond to the assumption of equal variances that was discussed in the motivational example, in which case $H_0 : \delta = 0$, where $\delta = \sigma^2_{boys} - \sigma^2_{girls}$.

The researcher has some prior beliefs about the plausibility of this assumption, based on the background information that he has about the particular situation he is dealing with. Since research never takes place in complete isolation from all previous research or substantive theory, the researcher will always have some background knowledge that is relevant for the particular context that he is in. If the researcher assigns either a probability of 0 or 1 to the model assumption being true before observing the data, he will consider statistically testing this hypothesis to be redundant, since he believes it cannot provide him with useful evidence about the model assumption. Thus, if researcher $j$ tests $H_0$, we have to assume that

$$0 < P_j(H_0) < 1. \tag{5.1}$$

Let us also assume that the researcher applies a null hypothesis test to the data $\boldsymbol{X}$, and that he contrasts $H_0$ with an alternative simple hypothesis $H_i$ (e.g., specifying the variances to differ by a specific amount). This procedure results either in a significant or a nonsignificant test statistic. For now we will assume that the researcher follows the guidelines of the liberal approach to NHST. Thus, a significant test statistic results in the researcher rejecting $H_0$ – the event of which is denoted by $R$ – and a nonsignificant value means that $H_0$ is accepted – denoted by $\neg R$.

For the test statistic to provide some form of justification for accepting

or rejecting[2] $H_0$ over $H_i$, it must also be the case that

$$P(R|H_0) < P(R|H_i). \tag{5.2}$$

That is, only if the probability of obtaining a significant test statistic under $H_i$ is larger than under $H_0$ can the null hypothesis test successfully inform us whether the data provide support for $H_0$ over $H_i$ or vice versa. Note that these probabilities do not depend on the subjective beliefs of our researcher, since both hypotheses are simple and $P(R|H_0)$ and $P(R|H_i)$ follow from the properties of the test statistic that is used. From Equation 5.2 it follows that

$$P(\neg R|H_0) = 1 - P(R|H_0) > P(\neg R|H_i) = 1 - P(R|H_i). \tag{5.3}$$

Thus, a failure to reject $H_0$ is more likely under $H_0$ than under $H_i$.

For convenience, let us assume that Equation 5.2 holds for all possible simple alternatives of $H_0$ (which are all mutually exclusive and which as a set together with $H_0$ are exhaustive),

$$P(R|H_0) < P(R|H_i), \text{for all } i. \tag{5.4}$$

Equation 5.4 generally holds for NHST-based tests for model assumptions, such as Levene's test for equality of variances (Levene, 1960). Let us denote the composite hypothesis that is the complement of $H_0$ by $\neg H_0$. Because the complement incorporates all possible alternatives to $H_0$, $\neg H_0$ is also known as the 'catch-all' hypothesis (Fitelson, 2006; 2007).

Our assessment of the probability of obtaining a nonsignificant test statistic under the catch-all hypothesis depends on how plausible we consider each of the possible alternatives to $H_0$ to be. That is,

$$\beta_j = P_j(\neg R|\neg H_0) = \frac{\sum_i P(\neg R|H_i)P_j(H_i)}{\sum_i P_j(H_i)}, \tag{5.5}$$

where $\beta_j$ denotes person $j$'s assessment of the probability of a Type II error. Thus, $\beta_j$ depends on the person that evaluates it, since $P(R|H_i)$ may differ

[2]Or for preferring $H_0$ over $H_i$ to some degree if one uses the continuous approach.

for different $H_i$s and since persons may differ with respect to their values for each $P_j(H_i)$. The power to detect a violation of the model assumption under the catch-all hypothesis thus cannot be assessed without considering the prior probability of each of the possible alternatives to $H_0$.

Equation 5.4 implies that

$$\alpha = P(R|H_0) < 1 - \beta_j. \tag{5.6}$$

Equation 5.6 informs us that the power of the test to detect a violation of the model assumption is larger than the probability of a Type I error given the truth of $H_0$. From Equation 5.6 it also follows that

$$P(\neg R|H_0) = 1 - P(R|H_0) > 1 - P_j(R|\neg H_0) = P_j(\neg R|\neg H_0). \tag{5.7}$$

Equation 5.7 informs us that a nonrejection is more likely under $H_0$ than under $\neg H_0$. Thus, obtaining a nonsignificant test statistic should increase our assessment of the plausibility of $H_0$,

$$P_j(H_0|\neg R) > P_j(H_0). \tag{5.8}$$

Hence, a null hypothesis test can indeed provide some degree of confirmation for the model assumption it evaluates. However, based on Equation 5.8 alone, we do not know *how plausible* $H_0$ is after a failure to reject it, nor do we know *how much more plausible* it has become due to this nonrejection.

The degree to which $H_0$ has become more plausible after having obtained a nonsignificant test statistic can be determined (Kass & Raftery, 1995; Trafimow, 2003) by means of

$$\frac{P_j(H_0|\neg R)}{P_j(\neg H_0|\neg R)} = \frac{P_j(H_0)}{P_j(\neg H_0)} \frac{P(\neg R|H_0)}{P_j(\neg R|\neg H_0)} = \frac{P_j(H_0)}{P_j(\neg H_0)} \frac{1-\alpha}{\beta_j}. \tag{5.9}$$

That is, the odds of $H_0$ versus $\neg H_0$ increase by a factor $\frac{1-\alpha}{\beta_j}$ after having obtained a nonsignificant test statistic, and this ratio corresponds to the likelihood ratio $\frac{P(\neg R|H_0)}{P_j(\neg R|\neg H_0)}$. Since $\alpha$ is the significance level, its assessment does not depend on the person that assesses it. However, Equation 5.5

shows that $\beta_j$ – our assessment of how likely it is to obtain a nonsignificant test statistic under $\neg H_0$ – depends on how plausible we consider each of the simple hypotheses incorporated in $\neg H_0$ to be. In the context of our motivational example, this means that our assessment of the power of the test depends on what values for $\sigma^2_{boys}$ and $\sigma^2_{girls}$ we consider to be plausible. If we expect a large difference between the two variances, we would expect the testing procedure to have a higher power to detect these differences than if we expect a small difference.

Thus, our assessment of the extent to which $H_0$ has become more plausible depends on our subjective assessment of the power of the null hypothesis test. Only when two simple hypotheses are contrasted will $\beta_j$ not depend on the person assessing it. However, in the context of model assumptions the alternative hypothesis has to be the catch-all hypothesis, which is by definition composite. If instead of contrasting $H_0$ with $\neg H_0$ we decide to contrast it with a simple alternative hypothesis $H_i$ we may be able to determine $\beta_j$ for that comparison objectively, but this does not inform us how much more plausible $H_0$ has become. That is, determining that $H_0$ is much more plausible than a simple alternative $H_i$ still does not tell us to what extent $H_0$ itself has become more plausible, since we have not considered all possible alternatives to $H_0$. To determine how plausible $H_0$ is after having obtained a nonsignificant result, we thus cannot avoid relying on a subjective assessment of the power of the test based on what we consider to be plausible alternatives to $H_0$.

The estimated effect size might seem to be a useful starting point that can help assess the actual (i.e., objective) probability of a Type II error, $\beta$ (Cohen, 1988; Thomas, 1997). However, the actual effect size (i.e., based on the population values of the parameters) rather than the estimated effect size (based on the sample values) is needed to calculate $\beta$, and these two effect sizes may differ due to sampling error. Thus, the actual effect size needs to be assessed before $\beta$ can be calculated, and this assessment of the effect size depends on the prior plausibility of each of the possible alternatives to $H_0$.

Consider an extreme hypothetical example, where we know exactly which alternative hypothesis is true, and hence we can assign a prior prob-

ability of 1 to that specific alternative. In that case, it would be irrational to base our assessment of the effect size and of $\beta$ on the data rather than on our knowledge of the values of the parameters in the population, since the latter approach guarantees that we obtain the true values while the former approach is subject to sampling error.

We can also consider a less extreme example, where in some study the estimated effect size was much larger than expected. For example, imagine that the researcher finds that the variance of the speed of the boys is ten times that of the girls, which is a much larger difference than he expected. Then, based on his background knowledge it may be reasonable to assume that the actual effect size is smaller than the observed one, implying that the estimated power is probably higher than the actual power. Consequently, our prior knowledge has an influence on the assessment of $\beta$. Even if we simply use the estimated effect size as our best guess, we are still assuming prior knowledge that each possible alternative to $H_0$ is equally probable (Jeffreys, 1961). Thus, assessing the likelihood ratio $\frac{1-\alpha}{\beta_j}$ requires one to take the prior probabilities of the hypotheses into account.

By combining Equation 5.9 with the fact that $P_j(H_0) = 1 - P_j(\neg H_0)$, we can obtain the plausibility of $H_0$ after having observed a nonsignificant result through

$$\begin{aligned} P_j(H_0|\neg R) &= \frac{P_j(H_0)P(\neg R|H_0)}{P_j(H_0)P(\neg R|H_0)+P_j(\neg H_0)P(\neg R|\neg H_0)} \\ &= P_j(H_0)\frac{1-\alpha}{\beta_j+P_j(H_0)(1-\alpha-\beta_j)}. \end{aligned} \tag{5.10}$$

Equation 5.10 shows that our conclusion about the plausibility of $H_0$ should depend on our prior assessment of its plausibility. It also shows that the degree to which $H_0$ has become more plausible depends on our assessment of the power of the test, which Equation 5.5 showed to depend on the prior plausibility of $H_0$ as well. Thus, it is not possible to assess the degree to which the data support $H_0$ through NHST alone, and the continuous approach to NHST cannot succeed if it does not take the prior plausibility into account.

Equation 5.10 also illustrates why the liberal approach to NHST cannot provide us with defensible guidelines for accepting or rejecting $H_0$. Since

the liberal approach does not take the prior plausibility of $H_0$ into account, it has to make a decision about the plausibility of $H_0$ based on the $p$-value alone (Trafimow, 2003). However, the $p$-value only tells us whether the data are consistent with the assumption being true, not whether this assumption is actually likely to be true. That is, a $p$-value only informs us of the probability of obtaining data at least as extreme as the data that were actually obtained conditional on the truth of the null hypothesis. It does not represent the probability of that hypothesis being true given the data (Wagenmakers, 2007). That is, $P(H_0|\neg R) \neq P(\neg R|H_0)$.

The fact that in practice the $p$-value is often mistaken to represent the probability that the null hypothesis is true (see e.g. Guilford, 1978; Gigerenzer, 1993; Nickerson, 2000; Wagenmakers & Grünwald, 2005; Wagenmakers, 2007) already suggests that it is this probability that we are often interested in (Gigerenzer, 1993). However, because they do not address the prior plausibility of the assumption, both the liberal and the continuous approach to NHST are unable to inform the user how plausible it is that the assumption is true after having taken the data into consideration.

Some proponents of NHST might argue that we still should avoid the subjective influence introduced by including the prior plausibility in our assessment of model assumptions, and that an objective decision rule based on the liberal approach is still acceptable. They might state that we simply have to accept uncertainty about our decision about the model assumption: If we repeatedly use NHST to evaluate model assumptions we will be wrong in a certain proportion of times, and this is something that simply cannot be avoided. But the problem is that the proportion of times we can expect to be wrong if we simply accept $H_0$ when the test statistic is not significant also depends on the prior probability of $H_0$, as Equation 5.10 shows. As mentioned before, if our model assumption cannot possibly be true, all failures to reject $H_0$ are Type II errors, and the decision to accept $H_0$ will be wrong 100% of the time. Thus, this uncertainty about the plausibility of the model assumption cannot be assessed without also assessing the prior plausibility.

If the truth of the model assumption is unknown to the person evaluating it, the proportion of times in which he incorrectly accepts $H_0$ also

Table 5.1: Proportion of acceptances of $H_0$ based on a null hypothesis test that person $j$ can expect to be wrong, for varying levels of power and prior plausibility ($\alpha = .05$)

| | $1 - \beta_j$ | | | | |
|---|---|---|---|---|---|
| $P_j(H_0)$ | .20 | .50 | .80 | .90 | .99 |
| .1 | .88 | .83 | .65 | .49 | .09 |
| .2 | .77 | .68 | .46 | .30 | .04 |
| .5 | .46 | .34 | .17 | .10 | .01 |
| .8 | .17 | .12 | .05 | .03 | .00 |
| .9 | .09 | .06 | .02 | .01 | .00 |

depends on the power of the statistical test. Proponents of NHST often argue in favor of doing a power analysis (see e.g. Cohen, 1988; Neyman, 1950). However, even if we disregard the fact that the power cannot be assessed without addressing the prior plausibility of the alternatives to $H_0$, establishing that the test has a high power does not warrant the conclusion that the proportion of incorrect decisions about our model assumption following the decision rule of the liberal approach is low.

The impact of the prior probability on the proportion of incorrect acceptances of $H_0$ based on a null hypothesis test ($\alpha = .05$) is illustrated in Table 5.1. For each particular combination of the power and the prior plausibility, Table 5.1 shows the probability with which person $j$ can expect the acceptance of $H_0$ based on a nonsignificant test statistic to be incorrect. That is, if based on a null hypothesis test he accepts $H_0$ as true while based on his prior knowledge he would conclude that the posterior plausibility of $H_0$ is .5, he can expect such a decision to accept $H_0$ to be wrong half of the time. The probability of incorrectly accepting $H_0$ is obtained using Equation 5.10, where $P_j(\neg H_0|\neg R)$ is obtained using $1 - P_j(H_0|\neg R)$.

Table 5.1 shows that even with a power as high as .90, person $j$ can determine that he will incorrectly accept $H_0$ in about 49% of times if a priori he considers the assumption to be implausible ($P_j(H_0) = .1$). Thus,

while power analysis is important in evaluating the support that the model assumption receives, Table 5.1 illustrates that a high power alone is not sufficient to result in convincing claims that the assumption is plausible enough to be accepted (barring theoretical cases where the power is 1). Without assessing the prior plausibility, the plausibility of the assumption after having observed the data can have any value between 0 and 1 regardless of the $p$-value that is obtained.

It may be emphasized that many of the most common model assumptions seem to have a very low a priori plausibility. Model assumptions quite often specify certain parameters to be zero (e.g., equality of variances in ANOVA, absence of multicollinearity between predictors in a linear regression model; Neter, Kutner, Nachtsheim & Wasserman, 1996). Of all the possible values that a parameter could take on, why should we find it plausible that a parameter has a value of precisely 0? Why would it be plausible that the effect that one variable has on the other is perfectly linear, or that an attribute has a normal distribution in the population? Why would we expect boys to show exactly the same variance in cognitive processing speed as girls?

Often model assumptions are chosen not because they are deemed plausible, but because of their mathematical convenience or usefulness to develop a statistical model. However, if there is no substantive theory that backs up these model assumptions with convincing arguments, there is little reason to assume that the model assumptions actually hold, and assigning a potentially very low prior plausibility to these assumptions may be the only reasonable response. This relates to Box's famous quote that "all models are wrong" (1987). Our models try to simplify reality with the goal of representing it in a convenient and useful way, but because of this simplification those models often cannot completely capture the vast complexity of the reality they try to represent. As such, the idea that they are completely correct may in many cases be highly implausible.

As a response to the alleged implausibility of statistical model assumptions that specify precise values for certain parameters, it is common practice to consider the robustness of the model: If the model is robust, inferences made using the model might still be approximately correct if the vio-

lations of the assumptions are not too severe. For many statistical models, much research with respect to robustness has been done. This definitely helps us to move away from implausible model assumptions that specify point values for parameters that could take on infinitely many other values. However, the question of how we can determine to what extent we can trust our model assumptions remains equally important here: If we want to know to what extent we can trust inferences based on the model, the model assumptions should still not be violated too severely. Hence, the evaluation of model assumptions remains relevant for all statistical models, even if some of these models are relatively robust. The questions 'Is the model assumption violated?' and 'Is the model assumption violated beyond an acceptable limit?' do not really differ in this regard: Both assumptions can be formulated as null hypotheses that can be evaluated using NHST based tests. The main change would be that the null hypothesis takes the form of an 'about equal' hypothesis (Hoijtink, 2012). Thus, in the context of the example we could test $H_0 : \sigma^2_{boys} \approx \sigma^2_{girls}$ rather than $H_0 : \sigma^2_{boys} = \sigma^2_{girls}$ if we know that the $t$-test is robust against violations of the assumption of equal variances, since we will still trust our model if the variances are not too different.

The main benefit of taking the robustness of the model into account in the formulation of the null hypothesis is that about equal hypotheses are more lenient with regard to the parameter values they allow, and therefore $H_0$ may receive a higher prior plausibility if it corresponds to an about equal hypothesis rather than a strict equality. That is, in many cases it is more plausible that a parameter value is 'about' equal to a certain value than that it is exactly equal to that value. But regardless of whether we formulate our model assumptions in the form of a regular null hypothesis or an about equal null hypothesis, we want to assess the plausibility of the assumptions that we have to make. Thus, regardless of the precise specification of the null hypothesis, the prior plausibility of that assumption needs to be assessed.

## 5.4 A Case Study: Evaluating Latent Monotonicity

In this section, the importance of taking the prior plausibility of model assumptions into account is illustrated using an example from item response theory (IRT). In IRT, one of the assumptions shared by almost all models is that of latent monotonicity. For example, in nonparametric IRT for dichotomous items, it is one of the assumptions defining the monotone homogeneity model (Mokken, 1971). Latent monotonicity states that for each of the items in a test the probability of observing a positive response is nondecreasing in the latent variable, and as such the assumption captures the idea that the items measure the ability or trait of interest (Junker & Sijtsma, 2000). That is, if latent monotonicity does not hold, for some range(s) of the latent variable subjects show worse performance on the item as the value on the latent variable increases. In that case, an increase in ability does not always result in a better performance on the item and the item does not measure the ability well. Because of this, latent monotonicity is an assumption that has substantive importance.

Let $\theta$ denote the latent variable, and let $X_i$ with realization $x_i$ denote the response to an item $i$ for which latent monotonicity is evaluated. Latent monotonicity for a test containing a set of items corresponds to

$$P(X_i = 1|\theta_a) \leq P(X_i = 1|\theta_b), \quad \text{whenever } \theta_a < \theta_b, \text{ and for all } i. \quad (5.11)$$

### 5.4.1 NHST-Based Evaluation of Latent Monotonicity

The evaluation of latent monotonicity is complicated by the fact that the latent variable is unobservable, so that Equation 5.11 cannot be evaluated directly. Instead, researchers attempting to evaluate latent monotonicity have to resort to evaluating some set of observable consequences that follow from Equation 5.11. There are various observable consequences that can be considered (see e.g. Rosenbaum, 1984; Tijmstra, Hessen, Van der Heijden & Sijtsma, 2013), which usually take the form of a set of constraints on a manifest score that is constructed using the items in the test (e.g., the

sumscore on a subset of the items). A variety of statistical procedures have been developed that evaluate these manifest properties (Rosenbaum, 1984; 1987b; Ramsay, 1991; Abrahamowicz & Ramsay, 1992; Scheiblechner, 1995; Junker & Sijtsma, 2000; Karabatsos & Sheu, 2004; Tijmstra et al., 2013).

Most of these testing procedures are based on the NHST framework, and proceed by testing whether the data are unlikely to be observed if the manifest properties being evaluated are true. If the data show too much deviation from what is expected given the manifest properties, a significant test statistic is obtained. As a result, the manifest property is rejected and by implication latent monotonicity is also rejected. This way, these procedures aim to provide a statistical test for latent monotonicity that can be used to determine whether the application of an IRT model might be justified.

These testing procedures for latent monotonicity operate in line with NHST: A significant test result means that we should be worried about applying the IRT model, and a nonsignificant test result is taken to suggest that we are relatively safe when we apply the model, especially if the power of the test was high. In addition to considerations about power, the question also arises whether the manifest score sufficiently 'covers' the full range of the latent variable. If this is not the case, then the manifest score does not give us enough information about some ranges on the latent variable to effectively assess latent monotonicity. That is, the manifest property might hold even if latent monotonicity is violated. Hence, even if it is established that the manifest property holds we still cannot be completely certain that latent monotonicity holds. By including a sufficient number of items in the manifest score, one hopes that the manifest score sufficiently covers the whole range of the latent variable to adequately assess latent monotonicity through the manifest properties it implies.

Tijmstra et al. (2013) discuss an example of a test for latent monotonicity based on NHST, where the null hypothesis corresponds to manifest monotonicity, which is an observable consequence of latent monotonicity. They showed that for realistic situations, this test may achieve a power of .80 or even .90 to detect a moderate violation of manifest monotonicity. Given this power, the procedure can definitely be considered to provide a

'critical test' (Mayo, 1996) for manifest monotonicity: If there is a serious violation of manifest monotonicity, the violation has a high probability of being detected by the procedure. Thus, since a nonrejection is not likely if manifest monotonicity is violated, a nonrejection makes it more plausible that latent monotonicity holds. Procedures such as these are able to provide us with useful information about the consistency of the data with the property they evaluate, and when a high power is achieved the procedure can make the assumption much more plausible.

However, NHST-based procedures such as these do not tell us *how* plausible the assumption of latent monotonicity is after having failed to reject the observable consequence they tested. Neither do these procedures inform us exactly *to what extent* the data support this observable consequence or latent monotonicity itself. Even though a nonrejection provides some support for the model assumption, the exact amount of support remains unclear. As Table 5.1 showed, even if we know that the test was powerful, we still do not know how plausible the assumption it tested is and hence to what extent we can trust inferences based on the IRT model.

### 5.4.2 Evaluating Latent Monotonicity Using Bayes Factors

To assess the amount of support that the data provide in favor of (or against) manifest monotonicity, Tijmstra, Hessen, Van der Heijden, Sijtsma & Hoijtink (submitted) proposed a Bayesian procedure. This procedure deviates from NHST-based procedures in that it allows users to take prior beliefs about the plausibility of manifest monotonicity into account. Through the use of Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995), the procedure quantifies the support that one hypothesis receives from the data relative to another competing hypothesis. By contrasting manifest monotonicity with its complement, a Bayes factor can be obtained that quantifies the extent to which the data favor one hypothesis over the other.

The Bayes factor can be translated into a decision rule about when manifest monotonicity should be accepted (Tijmstra et al., submitted). However, Bayes factors go beyond the NHST-based dichotomy of accepting versus rejecting manifest monotonicity, since they inform the user precisely

how strong the support is that manifest monotonicity receives from the data. A Bayes factor thus provides the user with more information than traditional NHST-based procedures do, and a more informed decision can be made about whether the model assumption should be accepted.

While the Bayes factor provides information about the *relative* support manifest monotonicity receives over its complement, it does not automatically provide us with an assessment of the posterior plausibility of manifest monotonicity. Since the Bayes factor is the ratio by which one hypothesis has become more likely than another after having observed the data, it can be used to obtain the posterior plausibility if the prior plausibility is specified, similar to Equation 5.10 (see also Kass & Raftery, 1995).

The Bayes factor resembles the likelihood ratio $\frac{P(\neg R|H_0)}{P_j(\neg R|\neg H_0)}$ that was described in Equation 5.9, but instead of evaluating the likelihood of obtaining a nonsignificant test statistic ($\neg R$) it evaluates the likelihood of obtaining the observed distribution of the item responses conditional on the manifest score. Let $\boldsymbol{p}$ denote the vector containing the proportion of positive scores conditional on the manifest score and let $MM$ denote manifest monotonicity, then the Bayes factor contrasting manifest monotonicity with its complement $\neg MM$ corresponds to

$$BF_{MM,\neg MM} = \frac{P_j(\boldsymbol{p}|MM)}{P_j(\boldsymbol{p}|\neg MM)}, \tag{5.12}$$

and the posterior probability of $MM$ is obtained through

$$\begin{aligned} P_j(MM|\boldsymbol{p}) &= P_j(MM)\frac{BF_{MM,\neg MM}}{1+BF_{MM,\neg MM}} \\ &= P_j(MM)\frac{P_j(\boldsymbol{p}|MM)}{P_j(\boldsymbol{p}|\neg MM)+P_j(MM)(P_j(\boldsymbol{p}|MM)-P_j(S|\neg MM))}. \end{aligned} \tag{5.13}$$

Using Equation 5.13, the procedure proposed by Tijmstra et al. (submitted) can be used to assess the posterior plausibility of manifest monotonicity. However, this leaves us with two problems: the inferential step in moving from manifest monotonicity to latent monotonicity, and the specification of the prior plausibility $P_j(MM)$.

### 5.4.3 Inferences from the Manifest to the Latent Level

Since latent monotonicity is evaluated through its observable consequences, only an absolute or definite falsification of such a consequence would give conclusive information about the truth of latent monotonicity, while a definite confirmation or verification of that consequence would not guarantee that latent monotonicity holds. However, absolute falsification is unachievable through inherently probabilistic statistical procedures, regardless of whether these are Bayesian or based on NHST. Because of this, both evidence in favor of the assumption (e.g., a nonsignificant test statistic or a Bayes factor larger than 1) or against it (a significant result or a Bayes factor lower than 1) cannot result in certainty about the model assumption, and the inferential step in moving from the manifest level to the latent level has to deal with probabilities rather than certainties.

The inferential step for assessing the support in favor of latent monotonicity is complicated by the probabilistic relation between latent monotonicity and the manifest properties it implies. Latent monotonicity (denoted by $LM$) does not hold in all cases where these manifest properties hold, and for example assessing the probability $P(LM|MM)$ of latent monotonicity being true given that manifest monotonicity holds is not straightforward. Assessing $P(LM|MM)$ requires determining how plausible it is that the item response function is monotone given that manifest monotonicity holds. However, there are an infinite number of item response functions $P(X_i = 1|\theta)$ that are consistent with the set of probabilities conditional on the manifest score $M$, $P(X_i = 1|M)$, and some of these item response functions show a violation of latent monotonicity while others do not.

However, as the number of items included in the manifest score increases, it becomes more and more plausible that properties at the manifest level translate to properties at the latent level. That is, the possibility that the manifest properties hold while violations of latent monotonicity are small or absent is realistic, provided that the items succeed in accurately covering a large enough range of the latent variable (but for a cautionary example see Junker & Sijtsma, 2000). It could be argued that if the items

sufficiently cover the range of the latent variable, a confirmation of the manifest properties is just as relevant for our evaluation of latent monotonicity as a falsification of the manifest properties.

Since the support of latent monotonicity through its observable consequences depends on whether the items sufficiently cover the range of the latent variable, it is crucially important for the evaluation of latent monotonicity that a large set of items with varying scale locations is used. This requirement holds for both NHST and Bayesian alternatives. That is, the uncertainty that arises in moving from the manifest properties to the latent level is also an issue for NHST-based approaches, since these approaches have to assume that a nonrejection of the manifest property informs us about the plausibility of the latent property. If NHST-based approaches do not claim that a nonrejection at the manifest level provides support for latent monotonicity, then it is unclear why we are any better off after not rejecting the manifest property than we were before we tested it, and the testing procedure becomes useless.

The apparent appeal of the NHST-based approaches may be that for the evaluation of latent monotonicity, an absolute falsification of one of its observable consequences is more informative than a verification of those properties: Only when the observable properties are is falsified can we draw a decisive conclusion about latent monotonicity. However, as Equation 5.7 illustrates, the support that a nonsignificant or a significant test statistic provides for $H_0$ is not absolute. Just as a Type II error may have occurred when we do not reject $H_0$, a Type I error could have occurred when we reject $H_0$, and hence no absolute conclusions can be drawn in either case. Thus, falsification in the NHST framework is in principle no less decisive than verification and remains probabilistic. Whether we can assume latent monotonicity to be plausible after having used a NHST-based procedure thus depends on how plausible latent monotonicity was in the first place.

### 5.4.4 Specifying a Prior for Latent Monotonicity

If we assume for the sake of argument that we can resolve the issue of translating support for certain properties at the manifest level into degrees

of support for latent monotonicity, the question still remains how we would assess the prior plausibility of latent monotonicity and the manifest properties it implies. Determining in advance whether it is plausible that latent monotonicity holds for a particular item is difficult and may depend to a large extent on substantive information about the item and the test that it is a part of.

Since latent monotonicity only puts a qualitative constraint on the function $P(X_i = 1|\theta)$ instead of specifying a precise parametric shape for this function, and since items are usually designed with the explicit purpose of measuring a specific ability or trait, a case could be made that latent monotonicity usually has at least some plausibility. If we consider a well-established set of items that has thoroughly been evaluated by experts and has already been used before successfully, latent monotonicity could perhaps even be considered to be highly plausible and one may consider assigning a high value to the prior probability of latent monotonicity being true. A high prior plausibility of latent monotonicity would then translate to a high prior plausibility for manifest monotonicity. Using this prior plausibility of manifest monotonicity, statistical tests such as the one proposed by Tijmstra et al. (submitted) could be used to assess the posterior plausibility.

If in evaluating manifest monotonicity one wants to assume a completely uninformed starting point (e.g., to avoid subjective biases), one could attempt to specify a prior that does not favor a monotone ordering of the item probabilities over other possible orderings (see also Tijmstra et al., submitted). However, the decision to use such an 'objective' prior is itself already a subjective decision not to take any background information into account and to assume that every ordering is equally likely (see also Jeffreys, 1961; Gillies, 2000). Thus, there seems to be no way of avoiding some degree of subjectivity in specifying the prior.

Translating these prior considerations about the plausibility of latent monotonicity to convincing probability values is an unavoidably subjective process. Ideally, the data will provide us with sufficient information to largely overrule the subjective influence of the prior, but there will definitely be situations where one person would consider the model assumption to be

warranted while another person will object to that conclusion. Especially in cases where the data are not overly informative about the model assumption it is important to back up claims about the prior that one has selected with substantive arguments for the (im-)plausibility of the model assumption in that particular case. Inconvenient and difficult as this may be, it is the only way to determine whether we can trust inferences made using any of the IRT models that assume latent monotonicity.

## 5.5 Conclusion

Evaluating the plausibility of model assumptions is crucial for justifying the use of a statistical model for making inferences. However, traditional NHST-based approaches to testing statistical model assumptions only provide ordinal information about the plausibility of the assumptions after having taken the data into account. Whereas NHST is normally employed with the aim of falsification in mind and generally avoids making statements about the plausibility of the null hypothesis, in order to justify using the statistical model to make inferences we are forced to move away from falsification and attempt to determine how plausible it is that the model assumptions hold. Here NHST runs into difficulties, since NHST-based procedures do not quantify the extent to which the data support the null hypothesis, nor do they assess the plausibility of the null hypothesis after having taken the data into consideration. Because of this, NHST-based procedures alone do not provide sufficient information to determine whether it is plausible that the model assumptions hold and whether we can trust the inferences that are made using the model.

A good NHST-based test may still have the potential of providing a critical test for the model assumption, provided that the test is applied in the right situation. That is, when the power of the procedure is high and $\alpha$ is low, obtaining a nonsignificant test statistic makes the model assumption much more plausible than it was before having observed the data. If the model assumption was already quite plausible in the first place, this might give us sufficient confidence in the assumption to apply the model. However,

if the power is low or the assumption was not plausible in the first place, a nonsignificant result provides us with insufficient reason to accept the model assumption, since it may not be very plausible.

NHST does not take prior plausibility into account, and it also cannot determine the actual power of the test without assessing the prior plausibility of the hypotheses. Thus, NHST-based approaches to evaluating model assumptions are insensitive to factors that should affect the conclusion that is drawn about the plausibility of the model assumption (see also Trafimow, 2003). Application of NHST without taking the prior plausibility into account may thus result in misleading conclusions about the plausibility of the model assumptions and about the validity of inferences made using the model.

By incorporating information about the prior plausibility of the model assumption and its alternatives in NHST-based testing procedures, the actual confirmatory power of such a procedure can be assessed. However, it may be more fruitful to abandon the idea that all the information about the model assumption is accurately captured by a dichotomized $p$-value (see also Cohen, 1994; Wagenmakers, 2007), and make use of all the information that is available in the data to assess the assumption. Plausibility is not a dichotomous concept, even if in the end we do want to make a decision about whether the assumption is plausible enough to apply the model. Recognizing that there are degrees of plausibility and degrees of support is important, and we have to acknowledge that there are situations in which we may not be sure if we are confident enough about the truth of our model assumptions to use the model. If we conclude that the assumption is not as plausible as we would have liked, we will have to be more cautious in using the model to make inferences, or we may conclude that we do not have enough confidence in the model assumptions and refrain from applying the model.

Prespecifying a general 'minimal level' of plausibility (i.e., a minimal value for the posterior plausibility) that is needed before we can safely apply the model would ignore that different situations call for different degrees of certainty about our inferences. Confirmatory analyses may call for higher levels of certainty than exploratory analyses, and high-stakes

testing situations may require even more certainty about the assumptions before we draw any conclusions. Thus, the choice for the required level of plausibility should depend on contextual factors. This also means that it is always possible to legitimately question whether an assumption was indeed plausible enough to warrant the conclusions that are drawn using the model. While the relevance of these contextual factors has been recognized within the NHST framework (Neyman & Pearson, 1967), it is usually only dealt with indirectly by adjusting the level of significance to control the Type I error rate. Adjusting the minimum required degree of plausibility would provide a more direct way of controlling the minimum degree of certainty that is desired.

The importance of taking prior plausibility into account when evaluating model assumptions is not necessarily restricted to NHST, but may be equally relevant for other methods of evaluating the plausibility of model assumptions. For example, if one uses graphical methods to determine whether it is plausible that the data are normally distributed, the extent to which we are confident that the assumption of normality holds after having seen that the distribution of the data resembles a normal distribution should likewise depend on whether it is actually plausible that the data come from a normal distribution. If we have reasons to suspect that this particular type of data usually comes from a skewed distribution (e.g., data on yearly income), we may not be fully confident that the data are indeed normally distributed, even if the graph does not show severe violations.

Having to deal with prior plausibility may complicate the way in which model assumptions are evaluated, but it is at the gain of being able to determine how much confidence we should have in the statistical inferences that we make using our models. Without assessing this prior plausibility, it will always remain uncertain whether a model can be trusted. Even worse, we will not have any information about the extent of this uncertainty either if we do not take the prior plausibility into account. The option of ignoring the prior plausibility of the model assumptions is thus not available if we ever want to argue that inferences based on statistical models have some degree of plausibility.

# Chapter 6

# Epilogue

This dissertation dealt with the evaluation of model assumptions in dichotomous item response theory (IRT), in particular the assumption of latent monotonicity and of invariant item ordering. These assumptions specify constraints on the item response function (IRF) $P(X_i = 1|\theta)$ for each item, and together with other assumptions define the IRT model. These other assumptions usually take the form of local independence and unidimensionality, which inform us that after taking the latent variable into consideration, no other variable can be invoked that helps explain the item responses and their interrelation. Since these two assumptions together do not yet inform us in what way the latent variable is related to the probability of observing a certain response, the assumptions of local independence and unidimensionality are not sufficient to obtain a measurement model.

To obtain a measurement model that is able to give us information about the ability or trait of persons based on their item responses, constraints have to be imposed on the IRFs. By including the assumption of latent monotonicity, an IRT model is obtained that secures an ordinal level of measurement for persons. The assumptions of local independence, unidimensionality and latent monotonicity together define the monotone homogeneity model (Mokken, 1971), a general IRT model that does not require the IRF to have a specific parametric shape. Because of its generality,

this model can be used in a wide range of applications, even in situations where the specific shape of the IRF differs for each item or is not precisely known.

Latent monotonicity tells us that the probability of obtaining a positive response increases (or at least does not decrease) as the latent variable increases. Thus, persons with a higher ability or trait value also have a higher probability of answering an item correctly or positively if latent monotonicity holds. Hence, latent monotonicity captures the idea that the items measure the latent variable.

Although the assumption of latent monotonicity is part of many IRT models and plays a crucial role in ensuring that the unweighted sumscore constitutes an ordinal scale, evaluating whether latent monotonicity holds is difficult. Because the latent variable by definition is unobservable, the shape of the IRF cannot be evaluated directly. Instead, latent monotonicity has to be evaluated based on its implications at the manifest level, which is the level of the item responses.

Both Chapter 2 and Chapter 3 focus on evaluating an observable consequence of latent monotonicity. If we assume that local independence holds, latent monotonicity for item $i$ implies that the probability $P(X_i = 1|Y = y)$ of obtaining a positive score on item $i$ is nondecreasing as the manifest score $Y$ (e.g., the unweighted restscore) increases. This property is called manifest monotonicity. As long as item $i$ is not included in the manifest score $Y$, manifest monotonicity over $Y$ has to hold if local independence and latent monotonicity hold. Since the restscore contains information from all the remaining items, it generally forms a more reliable ordinal estimator of the latent variable than a manifest score including fewer items (Junker & Sijtsma, 2000). Hence, using the restscore as the manifest score over which to evaluate manifest monotonicity is an attractive choice.

Chapter 2 showed that manifest monotonicity can be tested using the order-constrained statistical inference framework. The proposed procedure evaluates for individual items if the data are consistent with the order-constraints that are imposed by manifest monotonicity on the conditional item probabilities $P(X_i = 1|Y = y)$. Since there are many possible sets of values for these conditional probabilities that are consistent with manifest

monotonicity, this procedure makes use of the least favorable null distribution, where $P(X_i = 1|Y = 0) = \ldots = P(X_i = 1|Y = h)$. By making use of the least favorable null distribution, a $\bar{\chi}^2$-statistic is obtained for which the Type I error rate of the procedure is guaranteed not to exceed $\alpha$. Through simulation, the $p$-value corresponding to this $\bar{\chi}^2$-statistic can be approximated to any degree of precision, and a decision can be made about whether the data are consistent with manifest monotonicity. A simulation study was presented that showed that this procedure has high power to detect violations of manifest monotonicity under a variety of reasonable conditions.

Chapter 3 provided an alternative approach to evaluating manifest monotonicity, using the Bayes factor to determine whether manifest monotonicity is *supported* by the data. The Bayes factor quantifies the degree of support in the data in favor or against manifest monotonicity over a specified alternative. This way, manifest monotonicity can be contrasted with its complement (i.e., any ordering of $P(X_i = 1|Y = y), y = 0, \ldots, h$, is allowed except for the monotone ordering), and it can also be contrasted with more specific informed alternatives (e.g., only allowing for orderings of the conditional probabilities that are approximately monotone). Contrasting manifest monotonicity with its complement can be useful to determine whether there is general support for manifest monotonicity. On the other hand, contrasting manifest monotonicity with a more specific alternative can be useful to determine whether it is plausible that the conditional probabilities $P(X_i = 1|Y = y)$ do not just show an increasing trend (i.e., are essentially monotone), but that they really follow a strictly monotone order. For this reason, a two-step testing procedure was proposed in which both these comparisons are made. This way, users should be able to obtain a critical assessment of manifest monotonicity, with which they can determine whether there is general support in favor of manifest monotonicity, and whether it is plausible that the ordering is strictly monotone. A simulation study showed that this procedure successfully determines whether there is general support for manifest monotonicity, but that it is more difficult to find support for the claim that the ordering of $P(X_i = 1|Y = y)$ is strictly monotone rather than essentially monotone. However, for some

applications it may suffice to conclude that the orderings are essentially monotone (see also Van der Ark, 2005).

In addition to imposing the constraint of latent monotonicity on the IRFs, it may be attractive to restrict the IRFs in such a way that they do not intersect. This amounts to assuming an invariant item ordering, and this assumption can be attractive in a variety of applications, including DIF-analysis, adaptive testing and the analysis of aberrant response behavior (Sijtsma & Junker, 1996; Sijtsma & Molenaar, 2002). Additionally, having an invariant item ordering facilitates item interpretation. Chapter 4 proved that under the assumption of local independence, an invariant item ordering implies that the item-total regressions also are invariantly ordered. Thus, under the double monotonicity model (Mokken, 1971), the item-total regressions display an invariant ordering, which means that the testing procedure proposed by Karabatsos and Sheu (2004) can be used to test the double monotonicity model. Two measures of this invariant ordering of the item-total regressions were presented, which provide an indication of the extent to which the data are consistent with the assumption of invariant item ordering.

Chapter 5 focussed on an important philosophical question concerning the evaluation of model assumptions: How can we determine whether assumptions made by statistical models are plausible? It was argued that simply determining whether the data are consistent with the assumption being true by considering the $p$-value does not provide us with sufficient information to assess the plausibility of the model assumption. If we want to draw conclusions about the plausibility of the model and the extent to which we can trust statistical inferences made using the model, we also have to use our background knowledge about the plausibility of the model assumptions before we consider the data; that is, their prior plausibility. If we do not take this prior plausibility into account, we cannot assess the plausibility of the model assumptions after having observed the data. Additionally, without addressing the prior plausibility, we also cannot draw any conclusions about *how much more or less plausible* the model assumptions have become after having observed the data. The reason for this is that we cannot determine the power of the testing procedure without taking

into account which alternatives to $H_0$ – which corresponds to the model assumption – are plausible.

The conclusion that our assessment of the plausibility of a model assumption should not be based on the data alone has important implications for the procedures discussed in the other chapters, since it implies that users should be careful in basing their conclusions about whether latent monotonicity and invariant item ordering hold solely on the value of the calculated test statistic. Rather, they may have to take background information about the plausibility of the assumptions they are evaluating into account when determining what conclusions to draw based on the test statistics these procedures produce. As was described in Chapter 5, the Bayesian procedure proposed in Chapter 3 can readily be adapted to incorporate prior information about the plausibility of latent monotonicity. The procedure proposed by Karabatsos and Sheu (2004) that was discussed in Chapter 4 can likewise be adapted to incorporate prior knowledge about the plausibility of invariant item ordering, since it already requires the specification of a prior distribution for each probability $P(X_i = 1|Y = y)$.

One important message that has been emphasized throughout this dissertation is that the evaluation of model assumptions can only be done indirectly, by assessing observable consequences of these assumptions. Since the observable consequences may hold even when the assumptions themselves do not hold, concluding that it is plausible that a certain observable consequence is true does not automatically provide sufficient reason to conclude that the model assumption that implied that property also holds. This makes it especially difficult to determine to what extent the data provide support for the model assumptions, rather than just the manifest properties that these assumptions imply.

It should be emphasized that the proposed procedures for the evaluation of latent monotonicity and invariant item ordering depend on observable consequences which can only be derived given local independence. This means that when the tests proposed in this dissertation indicate that there has been a violation of latent monotonicity or invariant item ordering, this could also be a consequence of local independence being violated. Hence, before evaluating latent monotonicity and invariant item ordering

one should first assess local independence and unidimensionality using some of the available tests for these properties (see e.g. Lord, 1980; Rosenbaum, 1984; Hattie, 1985; Hambleton & Rovinelli, 1986; Holland & Rosenbaum, 1986; Stout, 1987; Gessaroli & De Champlain, 1996; Roussos, Stout & Marden, 1998). This way, it can be made plausible that for example a test result indicating a violation of manifest monotonicity is due to a violation of latent monotonicity, rather than a violation of local independence. This is important, since test constructors may have to take different types of action to deal with different types of violations. For example, one could imagine a situation where the proposed tests indicate that manifest monotonicity is violated. However, this result could be due to local independence being violated, perhaps because a second latent variable has to be taken into account to fully explain the association between the items. It could be that once the second latent variable is taken into account, the IRFs are in fact monotone over both latent variables.

The procedures proposed in this dissertation have been developed in the context of unidimensional IRT for dichotomous items. However, both latent monotonicity and invariant item ordering are assumptions that are also relevant when the items are polytomously scored (Hemker, Sijtsma, Molenaar & Junker, 1996; Ligtvoet, Van der Ark, Te Marvelde & Sijtsma, 2010). We expect that it should be possible to extend the procedures proposed in this dissertation to polytomous IRT. Likewise, it may be of interest to extend the current procedures to be applicable when multiple latent variable have to be taken into account. That is, the evaluation of latent monotonicity and invariant item ordering may also be relevant in the context of multidimensional IRT, where similar assumptions about the shape of the IRF can be made (see e.g. Reckase, 2009).

Furthermore, while the procedures that have been proposed aim at evaluating whether latent monotonicity and invariant item ordering hold, they do not provide guidelines about what users should do when they reject these assumptions or when they conclude that it is not plausible that these assumptions hold. While it is true that once one of the assumptions defining an IRT model is violated, that model is no longer strictly valid, this model might still be useful (Box & Draper, 1987). For example, it has been

shown that small violations of latent monotonicity need not have a severe impact on the ability of the total score to successfully order persons based on their ability (Van der Ark, 2005). This means that for some low-stakes testing applications essential monotonicity (as discussed in Chapter 3) may already be sufficient.

The extent to which an IRT model is still useful when some of its assumptions are violated depends on the extent and the severity of the violation, as well as the robustness of the model against such violations. This question of robustness has not been the focus of this dissertation, but it is nevertheless important. Using an IRT model when not all of its assumptions are met may be defensible if the model can be shown to be robust against the specific violation that is present. However, since the measurement properties implied by the model do not automatically hold when the model is formally invalid, the user is advised to treat inferences based on an IRT model for which the model assumptions do not hold with caution. The methods proposed in this dissertation can help users to assess whether IRT assumptions hold, and hence help to establish the measurement properties implied by the IRT model.

# References


Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57,* 5–27.

Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70,* 283–304.

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73,* 183–208.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces* (p. 424). New York: Wiley.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models.* New York, NY: Springer Science.

Dekovic, M. (2003). Aggressive and nonaggressive antisocial behavior in adolescence. *Psychological Reports, 93,* 610-616.

Earman, J. (1992). *Bayes or bust.* Cambridge: MIT Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Ellis, J. L., & Junker, B. W. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics, 25,* 1327–1343.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications.

Fisher, R. A. (1930). *Statistical methods for research workers* (3rd ed.). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B), 17*, 69–77.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* New York: Hafner.

Fisher, R. A. (1960). *The design of experiments* (7th ed.). Edinburgh: Oliver & Boyd.

Fitelson, B. (2006). Logical foundations of evidential support. *Philosophy of Science, 73,* 500–512.

Fitelson, B. (2007). Likelihoodism, Bayesianism, and relational confirmation. *Synthese, 156,* 473–489.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.

Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33,* 157–179.

Gibbons, J. D. (1985). *Nonparametric statistical inference* (2nd ed.). New York: Dekker.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as Intuitive Statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance. How probability changed science and every day life.* Cambridge, UK: Cambridge University Press.

Gillies, D. (2000). *Philosophical theories of probability.* London: Routledge.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–96). New York: Springer.

Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764.

Grayson, D. A. (1998). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Guilford, J. P. (1978). *Fundamental statistics in Psychology and Education* (6th ed.). New York: McGraw-Hill.

Hacking, I. (1976). *Logic of statistical inference.* Cambridge: Cambridge University Press.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287–302.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Nijhof.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*, 679–693.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Hessen, D. J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika*, *70*, 497–516.

Hoijtink, H. J. A. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton: CRC Press.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *The Annals of Statistics, 14,* 1523–1543.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach.* La Salle, IL: Open Court.

Hume, D. (1888). *Hume's treatise of human nature* (L. A. Selby Bigge Ed., originally published 1739–49). Oxford: Clarendon Press.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59,* 77–79.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response theory models. *The Annals of Statistics, 21*, 1359–1378.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24,* 65–81.

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*, 110–125.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kendall, M.G., & Babington Smith, B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, *10*, 275–287.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in Psychology: Comment on Trafimow (2003), *Psychological Review, 112*, 662–668.

Lehmann, E. L. (2006). The Fisher, Neyman-Pearson theories of hypothesis testing: One theory or two?, *Journal of the American Statistical Association, 88*, 1242–1249.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Stanford, CA: Stanford University Press.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70,* 578–595.

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing.* New York: Springer.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York: Springer.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago: The University of Chicago Press.

Michael, J. R. (1983). The stabilized probability plot, *Biometrika, 70*, 11-17.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Berlin: De Gruyter.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows.* Groningen: ProGAMMA.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York: Wiley.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53,* 99–138.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Ser. A, 236*, 333–380.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review, 25*, 7–22.

Neyman, J., & Pearson, E. S. (1967). *Joint statistical papers.* Berkeley: University of California Press.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Pearson, E. S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society (B), 17*, 204–207

Popper, K. R. (1959). *The logic of scientific discovery.* London: Unwin Hyman.

Popper, K. R. (1974), *Conjectures and refutations* (5th ed.). London: Routledge and Kegan Paul.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsey, F. P. (1926). Truth and probability. In Ramsey 1931, 156–198. Reprinted in H. E. Kyburg amd H. E. Smokler (Eds.), *Studies in Subjective Probability*, 1964, 61–92. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen and Lydiche.

Raven, J. C. (1956). *Progressive matrices, sets A, B, C, D and E.* London: H. K. Lewis.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425–435.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40,* 157–168.

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1–30.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover Publications.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika*, *60*, 281–304.

Scheiblechner, H. (2003). Nonparametric IRT: Testing the bi-isotonicity of isotonic probabilistic models (ISOP). *Psychometrika*, *68*, 79–96.

Sijtsma, K., & Junker, B. W., (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79–105.

Sijtsma, K., & Meijer, R.R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149–157.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: SAGE Publications.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order, and shape restrictions.* Hoboken, NJ: John Wiley & Sons, Inc.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Suppes, P. (2007). Where do Bayesian priors come from? *Synthese, 156,* 441–471.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191–199.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Thomas, L. (1997). Retrospective power analysis, *Conservation Biology, 11,* 276–280.

Tijmstra, J., Hessen, D. J., Heijden, P. G. M., & Sijtsma, K. (2011). Invariant ordering of item-total regressions. *Psychometrika*, *76*, 217–227.

Tijmstra, J., Hessen, D. J., Van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference, *Psychometrika, 78,* 83–97.

Tijmstra, J., Hessen, D. J., Van der Heijden, P. G. M., Sijtsma, K., & Hoijtink, H. J. A.. *Evaluating manifest monotonicity using Bayes factors.* Manuscript submitted for publication.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110,* 526-535.

Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical & Statistical Psychology*, *61*, 179–187.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of $p$ values, *Psychonomic Bulletin & Review, 14,* 779–804.

Wagenmakers, E.-J., & Grünwald, P. (2005). A Bayesian perspective on hypothesis testing. *Psychological Science, 17,* 641–642.

# Nederlandse Samenvatting: Het Evalueren van Modelassumpties in Item Respons Theorie

In de sociale wetenschappen wordt bij het meten van psychologische constructen vaak gebruik gemaakt van de item respons theorie. Met behulp van de statistische modellen uit de item respons theorie hopen onderzoekers informatie te verkrijgen over eigenschappen die niet direct te observeren zijn – zoals intelligentie of depressie – aan de hand van gedrag dat met deze eigenschap te maken heeft. De hoop is dat als personen een 'respons' leveren op een set van 'items' (bijvoorbeeld vragen in een intelligentietest), wij deze responsen kunnen gebruiken om informatie te verkrijgen over het psychologische construct dat wij willen meten (bijvoorbeeld intelligentie). Aan de hand van de statistische modellen uit de item respons theorie worden deze responsen op de items aan een (of meerdere) latente variabele(n) gerelateerd, in de hoop daarmee het psychologische construct waar de items betrekking op hebben te meten.

Zoals voor alle statistische modellen geldt zijn ook de modellen uit de item respons theorie gedefinieerd aan de hand van een aantal assumpties. Als we een dergelijk statistisch model gebruiken om aan de hand van observaties uitspraken te doen over een latent psychologisch construct, dan

zijn deze uitspraken alleen valide als het model zelf valide is – als alle assumpties die het model definiëren opgaan. Om het gebruik van deze item respons modellen te rechtvaardigen is het dus van belang dat er middelen zijn om na te gaan of deze modelassumpties inderdaad kloppen. In dit proefschrift worden verschillende methoden voorgesteld voor het evalueren van modelassumpties in de item respons theorie, met als doel de gebruiker ervan in staat te stellen kritisch naar deze assumpties te kijken en een beslissing te kunnen nemen over of deze assumpties aangenomen kunnen worden. Bij de meeste toepassingen van item respons theorie wordt gebruik gemaakt van dichotome data (bijvoorbeeld 'ja/nee'-vragen, of het correct of incorrect beantwoorden van een vraag op een test). De in dit proefschrift voorgestelde technieken richten zich op item respons modellen die bedoeld zijn voor het analyseren van dit soort dichotome data, maar kunnen met verder onderzoek in principe ook uitgebreid worden om modellen voor items met meer dan twee antwoordcategorieën te omvatten.

De modelassumptie die in hoofdstuk 2 en 3 centraal staat betreft latente monotonie. Deze assumptie vertelt ons iets over de relatie tussen de responsen op de items en de latente variabele: Hoe hoger iemands waarde op de latente variabele is, hoe groter de kans dat deze persoon een positieve respons levert op een item. Dit wil zeggen dat wij bijvoorbeeld verwachten dat iemand met een hoge intelligentie bij het oplossen van problemen als onderdeel van een intelligentietest meer kans heeft op het correct oplossen van deze problemen dan iemand met een lagere intelligentie. In feite vertelt deze assumptie ons dat het gedrag waar we naar kijken – de item responsen – op een eenduidige manier samenhangt met het construct wat we willen meten: Mensen die de te meten eigenschap in sterke mate bezitten zullen altijd een grotere kans (of op zijn minst geen kleinere kans) hebben op het leveren van een positieve respons op de items dan mensen die deze eigenschap in mindere mate bezitten. Latente monotonie is daarmee een belangrijke assumptie, omdat het garandeert dat wij de ongewogen totaalscore (bijvoorbeeld het aantal correcte items op een test) kunnen gebruiken om mensen te ordenen op de eigenschap die we meten.

Aangezien latente monotonie een uitspraak doet over de relatie tussen de latente variabele en de item responsen en de latente variabele zelf niet

geobserveerd kan worden, kan deze assumptie alleen geëvalueerd worden door te kijken naar observeerbare consequenties van deze assumptie. Uit de assumptie van latente monotonie volgt (samen met de assumptie van locale onafhankelijkheid) een aantal van dit soort manifeste eigenschappen die aan de hand van statistische methodes voor ieder individueel item geëvalueerd kunnen worden.

Een van de observeerbare consequenties van latente monotonie betreft manifeste monotonie. Manifeste monotonie vertelt ons dat als we mensen groeperen aan de hand van het totale aantal positieve scores op een set van items – de ongewogen somscore voor deze items –, de kans op het observeren van een positieve score op een specifiek item monotoon toeneemt als deze somscore toeneemt. Om het item waarvoor wij manifeste monotonie analyseren eerlijk te evalueren is het van belang dat dit item zelf niet meegenomen wordt in de somscore, omdat dit de resultaten kan vertekenen. Het ligt daarom voor de hand om manifeste monotonie te beschouwen over de restscore – de ongewogen somscore over alle items in de test behalve het item dat geëvalueerd wordt. Kort samengevat vertelt manifeste monotonie ons dat als iemand van de overige vragen meer goede antwoorden heeft gegeven dan iemand anders, de eerste persoon nooit een lagere kans kan hebben op het goed beantwoorden van het item waar we naar kijken.

In hoofdstuk 2 introduceren wij een statistische procedure die gebruikt kan worden om de assumptie van latente monotonie te evalueren door te testen of voor een specifiek item manifeste monotonie verworpen dient te worden. Deze procedure richt zich op het evalueren van de ongelijkheidsrestricties die door manifeste monotonie opgelegd worden op de kans op het goed beantwoorden van een item, conditioneel op de restscore. De voorgestelde procedure behandelt manifeste monotonie als een nulhypothese die gecontrasteerd kan worden met haar complement. Deze procedure levert de gebruiker een gewogen chi-kwadraat statistiek, waarvan de $p$-waarde door middel van simulaties tot de gewenste precisie benaderd kan worden. Aan de hand van deze $p$-waarde kan vervolgens een beslissing worden genomen over of de data dusdanig inconsistent zijn met manifeste monotonie dat latente monotonie verworpen dient te worden.

Ook hoofdstuk 3 richt zich op het evalueren van latente monotonie aan

de hand van manifeste monotonie. Echter, waar de procedure uit hoofdstuk 2 gebruikt maakt van frequentistische methodes om tot een test statistiek te komen, maakt de procedure die in hoofdstuk 3 voorgesteld wordt gebruik van Bayesiaanse methoden met als doel om de mate waarin de data manifeste monotonie ondersteunen te kwantificeren. Deze procedure richt zich niet zozeer op het verwerpen van manifeste (en daarmee latente) monotonie, maar probeert ook eventuele ondersteuning voor manifeste monotonie te evalueren. De mate van steun voor manifeste monotonie kan door middel van de Bayes factor gekwantificeerd worden, en deze Bayes factor kan bepaald worden met behulp van de in hoofdstuk 3 voorgestelde methode.

Hoofdstuk 4 richt zich op een andere assumptie uit de item respons theorie: invariante item ordening. Waar latente monotonie belangrijk is voor het ordenen van personen, betreft deze assumptie juist de ordening van de items. Een invariante item ordening vertelt ons dat de ordening van de items op basis van hun moeilijkheid niet afhangt van de waarde op de latente variabele, iets wat in een verscheidenheid aan toepassingen relevant is. In hoofdstuk 4 wordt aangetoond dat uit de assumptie van een invariante item ordening een observeerbare consequentie volgt die ingezet kan worden om deze assumptie te testen. We presenteren een statistische maat voor invariante item ordening die gebaseerd is op Kendalls $W$ en bespreken een Bayesiaanse procedure die gebruikt kan worden om te testen of er sprake is van een invariante item ordening.

De kern van hoofdstuk 5 richt zich niet op een specifieke modelassumptie, maar gaat in op de overkoepelende vraag welke statistische methoden ingezet kunnen worden om modelassumpties te evalueren. In dit hoofdstuk wordt geargumenteerd dat de wijze waarop nulhypothese-toetsen standaard worden ingezet om modelassumpties te beoordelen gebruikers niet voldoende informatie levert om te bepalen of de assumptie wel of niet aangenomen kan worden. Dit is problematisch, aangezien conclusies die aan de hand van een statistisch model getrokken worden alleen strikt valide zijn als alle modelassumpties opgaan. Om te bepalen hoeveel waarde we kunnen hechten aan de statistische inferenties die voortkomen uit een statistisch model is het noodzakelijk dat we de plausibiliteit van de assumpties die behoren bij dit model in overweging nemen. De wijze waarop dit gedaan

kan worden wordt gellustreerd aan de hand van de Bayesiaanse procedure voor het evalueren van latente monotonie die in hoofdstuk 3 is besproken.

# Acknowledgements

I would like to make use of this part of the dissertation to give thanks to the many people that have in some way contributed to the completion of this PhD project. First of all, I would like to thank my promotor Peter van der Heijden and co-promotor David Hessen for making this PhD project possible and for your support and useful feedback throughout the duration of this project. You have both given me a lot of freedom in giving shape to this dissertation, for which I am grateful. I would also like to thank Klaas Sijtsma, who joined the project as my second promotor shortly after it started. I have learned a lot from your expertise and dedicated feedback, and am thankful that you were always willing to provide long-distance support, no matter how busy your schedule may have been.

I would also like to thank Herbert Hoijtink, who sparked my interest in Bayesian statistics, is co-author of chapter 3 and provided useful feedback on an early version of chapter 5. My thanks also go to Janneke van Lith, who provided feedback on chapter 5, and whose support made it possible for me to combine working on my PhD with completing my research master in philosophy. Additionally, I would like to thank Maja Dekovic, who was generous enough to let me use her data to try out our statistical procedures. My thanks also go out to ETS and to Frank Rijmen in particular for inviting me to the ETS Summer Internship program, and to Patrick Kyllonen for welcoming me to his house during those two months.

I am especially grateful to all my colleagues at the M & S department, who provided a positive working environment that made my working place really feel like a second home. I have thoroughly enjoyed your company

and will look back with joy on the time we spent together, both inside and outside the office. I consider many of you to be my friends, and hope we will stay in touch. A special thanks goes out to the roommates that I have shared my office with over the years. I have enjoyed our on-topic and off-topic discussions, and highly appreciate your continued support. People who claim that the life of a PhD candidate is a lonely and solitary one have definitely not had roommates that are as great as you are.

My acknowledgements would definitely not be complete without mentioning the IOPS. Taking part in the biannual IOPS conferences has always been a positive experience for me, and I have also enjoyed the two years that I participated in the IOPS board as PhD representative. Most of all, I have valued the company of the IOPS PhD students, whose company made every national and international conference (and the trips we so conveniently combined with those conferences) always something to look forward to. Participating in these conferences and traveling together with you has definitely been one of the highlights of the past years.

A good life is a balanced life, and I am sure that this dissertation could not have been produced without the support I received from friends and family. Being able to spend time with you, take my mind of work or share whatever might have been troubling me with you has helped me enormously, and I very much appreciate that. A special word of thanks goes out to my mother, father and sister: you have always been there for me, and I could not have made it to where I am today without your support throughout the years.

Finally, I would like to thank Maria. Psychometrics brought us together, and we have often joked that there are now two Psychometrikas in my life. Your endless support, positive attitude and feedback have been of immense value to me.

# Curriculum Vitae

Jesper Tijmstra was born in 1985 on June 8 in Amersfoort, the Netherlands. He finished his preparatory university education at Stedelijk Gymnasium Johan van Oldenbarnevelt in 2003. Subsequently, he combined studying Psychology and Philosophy at Utrecht University, obtaining both bachelor degrees cum laude in 2006 and 2008, respectively. After obtaining his bachelor degree in Psychology, he followed the research master Methodology and Statistics in the Behavioural and Social Sciences at Utrecht University. He graduated cum laude for this master in 2008, and was jointly awarded the Statistics Netherlands prize for best master thesis.

In September 2008, Jesper started working as a PhD candidate at the Department of Methodology and Statistics of Utrecht University. During his PhD, he worked on a variety of topics both within and outside of item response theory, and he was involved in a large number of statistical consultations to students and researchers. In 2013, he was selected to participate in the ETS Summer Internship program, where he worked for two months on the development of software for multidimensional item response theory under the supervision of Frank Rijmen. A number of his work has appeared in journals such as Psychometrika.

In addition to doing research, he also acted as the representative of the PhD candidates of the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS), and was vice-chair of the PhD Council of the Faculty of Social and Behavioral Sciences at Utrecht University. He combined working on his PhD with a job as junior lecturer at Utrecht University. As a lecturer, he has acted as tutor of the research master Methodology and

Statistics, and has taught a variety of courses on both Statistics and Philosophy of Science. He combined his work at the Department of Methodology and Statistics with following the research master of Philosophy at Utrecht University, which he completed in 2013. Additionally, he is first author of two book chapters in *Onderzoeksmethoden*, and together with Hennie Boeije wrote the book *Wetenschapsfilosofie in de context van de sociale wetenschappen*, which is used in a variety of undergraduate courses.