Measurement and Prediction in Heterogeneous Populations

Maij-de Meij, A.M.

2008

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

Measurement and Prediction in Heterogeneous Populations

Printing: PrintPartners Ipskamp BV, Enschede
ISBN 978-90-86592-03-6

VRIJE UNIVERSITEIT

Measurement and Prediction in Heterogeneous Populations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Psychologie en Pedagogiek
op vrijdag 13 juni 2008 om 10.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Annette Maria Maij-de Meij

geboren te Amsterdam

promotoren: prof.dr. H. van der Flier

prof.dr. H. Kelderman

Voorwoord

November 2002 begon ik met mijn onderzoeksproject. Psychometrisch onderzoek, met mijn achtergrond als arbeids- en organisatie psycholoog. Het is een uitdaging geweest, waarbij uiteraard het een en ander mee maar ook tegen zat. Analyses met Lem die dagen konden duren, zelf leren programmeren, zware reviews, maar ook geaccepteerde artikelen en mooie resultaten. Het zijn onvergetelijke jaren geworden. Vele mensen hebben direct of indirect bijgedragen aan dit eindresultaat. Hen wil ik graag op deze plek bedanken.

Allereerst wil ik mijn promotoren Henk Kelderman en Henk van der Flier bedanken. Samen vormden jullie de juiste combinatie om de balans te vinden tussen psychometrie en psychologie. Bedankt voor alle tijd, commentaren, steun, vertrouwen en inspiratie. Daarnaast hebben jullie de betrokkenheid van NWO bij dit project bewerkstelligd (project 402-01-064). Ik wil NWO zeer bedanken voor het financieel mogelijk maken van dit project. De leden van de leescommissie, Klaas Sijtsma, Willem Heiser, Gunter Maris, Don Mellenbergh en Pim Cuijpers, wil ik graag bedanken voor de aandacht die zij aan dit proefschrift hebben geschonken.

Alle (ex-)collega's van de afdeling Arbeids- en Organisatie Psychologie hebben samen gezorgd voor een warme en gezellige werkomgeving. Ik dank Frank, Paul, Jan, Annebel, Barbara, Peter, Reinout, Ida, Anna, Dolly, Cathy, Wies, Edwin, Brigitte en alle anderen. Etentjes, borrels, dagelijkse lunches, en gewoon praatjes op de gang of bij de kopieer, jullie maakten het dagelijkse AIO leven des te aangenamer. Gert, een betere kamergenoot had ik me niet kunnen wensen, de tijd is voorbij gevlogen. Hanneke, Jacqueline en Annefloor; even bijkletsen of thee drinken, jullie waren altijd paraat.

Mensen van buiten mijn onderzoek hebben het regelmatig onduidelijk en ingewikkeld gevonden waar ik me al die jaren mee bezig hield. Jullie steun, interesse en afleiding zijn voor mij van grote waarde geweest. Barbara, vanaf het moment dat we elkaar leerden kennen ben je er. Niet dagelijks, maar toch altijd aanwezig. Ik vind het heel fijn dat je

bij deze belangrijke gebeurtenis mijn paranimf wilt zijn.

Mama, altijd heb je in me geloofd. Je hebt me geleerd dat met doorzettingsvermogen en vertrouwen alles mogelijk is. Ik vind het geweldig dat je mij bij de ceremonie als paranimf bijstaat. Papa, ik weet dat je van boven trots mee zult kijken. Arjen, fijn dat je erbij gekomen bent en zult blijven, en dank voor de prachtige omslag.

Bob en Iemkje, jullie hadden altijd aandacht voor waar ik me mee bezig hield, bedankt voor alle steun en interesse. Vooraf was het een plan van vier jaar. Het zijn er door de geboorte van twee prachtige zonen uiteindelijk vijf geworden. Koen en Lars, zonder dat jullie het weten hebben jullie bijgedragen aan het eindresultaat. Jullie zorgden voor afleiding, relativering en toch ook ontspanning. Mark, zonder jou zou dit me allemaal nooit gelukt zijn. Je bent mijn steun en toeverlaat, samen kunnen we alles aan. Toen, nu en voor altijd.

Middenbeemster, Januari 2008
Annette Maij-de Meij

Contents

Chapter 1

Introduction

In social and behavioral sciences many attributes cannot be observed directly. Measures of ability, personality, or attitude are obtained from observed responses gathered through, for example, tests and questionnaires. Statistical procedures are followed after subjects have responded to a set of items, resulting in a classification of subjects or quantification of their attribute. Latent variable models were introduced to account for observed patterns of association between the manifest, i.e. observed, variables (Lazarsfeld & Henry, 1968). A general modeling framework has been developed, relating the probability of an item response to an underlying latent variable, which can be either continuous or discrete. Based on the assumption of local independence, the association among manifest variables is explained by their dependence upon the latent variable (Lazarsfeld & Henry, 1968).

In item response theory (IRT), the latent variable is defined as a continuous latent trait. Measurement models have been developed to specify the probability of a response in terms of person and item parameters (e.g. Birnbaum, 1968; Guttman, 1950; Lord & Novick, 1968; Rasch, 1960). The latent trait describes the within-group heterogeneity, that is, quantitative individual differences. It is assumed that all subjects respond according to the same measurement model (Mellenbergh, 1989; Meredith, 1993). For an introduction to latent trait models the reader is referred to Embretson and Reise (2000), Hambleton, Swaminathan and Rogers (1991), or Van der Linden and Hambleton (1997).

Latent class models consider the underlying latent variable to be a discrete, categorical variable (Goodman, 1974a; 1974b; Lazarsfeld, 1950). Latent class analysis is applied to model the between-group heterogeneity, where the classes differ in a qualitative sense. The model assumes homogeneity within each latent class, that is, all subjects

within a latent class have equal response probabilities. An overview of the differences and similarities between latent trait and latent class models can be found in Heinen (1996) and Langeheine and Rost (1988).

Goodman (1975) combined the two latent variable models. Subjects were divided into a class that is considered "intrinsically scalable" according to the Guttman model, and an "unscalable" group of subjects where responses are independent of each other. The model was extended to include "demi-scale types" as well, where additional response patterns are allowed by the model (Goodman, 1975). Next, the development of latent class and latent trait models proceeded relatively independent. Kelderman and Macready (1990), Mislevy and Verhelst (1990), and Rost (1990; 1991) introduced the mixture IRT modeling framework. It was demonstrated that a given IRT model does not always hold for all subjects. A mixture IRT model identifies a number of unobserved groups of subjects, i.e. latent classes, for which different measurement models hold. In general, this means that the same measurement model holds within each latent class, but with different parameter estimates across latent classes. The model relaxes the assumption that the measurement model is invariant over subjects. In this way, mixture IRT models describe the heterogeneity in the population by modeling both quantitative and qualitative individual differences. This relates to the classification of heterogeneity made by Kelderman and Molenaar (2007).

Mixture IRT models specify the probability of a response of a subject given a subject's latent trait value and latent class membership. The models vary in degree of parsimony. They may contain item difficulty and/or discrimination parameters that may vary across latent classes and response categories. The mixture Rasch model is one of the most restricted mixture IRT models (Kelderman & Macready, 1990; Rost, 1990). The model contains item difficulty parameters which are allowed to vary across latent classes. The most general mixture IRT model used in this thesis is a mixture version of Bock's (1972) nominal response model (mNRM; Smit, Kelderman, & Van der Flier, 2003). The mNRM is a model for polytomous items, and allows the item response category difficulty and discrimination parameters to vary across latent classes. Exogenous variables may be added to mixture IRT models to improve estimation of the model parameters, and to aid in the interpretation of the latent classes (Smit, Kelderman, & Van der Flier, 1999; 2000). Assuming normality of the latent trait, maximum likelihood estimates of the model

parameters can be computed by means of the EM-algorithm (Dempster, Laird, & Rubin, 1977; Vermunt, 1997). Models with different numbers of latent classes are compared to determine the number of latent classes that provides the best fit to the data. Because models with different numbers of latent classes are not nested, information criteria, like AIC and BIC, are used to compare them. When the response behavior can be described by a simple IRT model, i.e. a one-class model, the items function the same way for all subjects. However, when there are qualitative differences in the responses of groups of subjects, more than one latent class should be identified.

In the remainder of this chapter, several areas of application for the mixture IRT modeling framework will be discussed. Throughout this thesis, quantitative differences described by the latent trait are conceived as differences in position of subjects on a metric scale. It is assumed that the same latent trait is measured within each latent class. We will look at qualitative individual differences from several perspectives. On the one hand, one may be primarily interested in differences in the way the measured attribute manifests itself in the item responses. On the other hand, qualitative differences in response patterns can be viewed as artifacts. Furthermore, we will investigate the use of mixture IRT models for DIF detection and prediction.

1.1 Differences in Kind and Degree

The main quality of the mixture IRT modeling framework is that it is sensitive to both quantitative and qualitative individual differences. The qualitative differences may be reflected in different patterns of response probabilities, while conditioning on the latent trait. Particularly relevant for the measurement of ability and aptitude, is the interpretation of the latent class variable, for example, in terms of describing differences in strategy use. Mislevy and Verhelst (1990) modeled item responses with a mixture version of the linear logistic test model. Latent classes corresponded to subjects employing different item-solving strategies (see also Gitomer & Yamamoto 1991). Mixture IRT models can be extended to include strategy shifts as well, like the latent response model (Rijkes & Kelderman, 2007; see also Yamamoto, 1995). Furthermore, the effect of test speededness on item responses can be studied (Bolt, Cohen, & Wollack, 2002; Yamamoto, 1995). The models may also be used to investigate cognitive development (e.g. Wilson,

1989) as well as qualitatively different patterns of change in longitudinal research (Meiser, Hien-Eggers, Rompe, & Rudinger, 1995).

For the measurement of personality attributes, a common distinction is between dimensions and categories (De Boeck, Wilson, & Acton, 2005), and traits versus types (Meehl, 1992). Gangestad and Snyder (1985) argued that discrete classes in personality exist, differentiating between high and low self-monitors. Von Davier and Rost (1997) studied the same data with a Hybrid model (Yamamoto, 1989). They agreed on the presence of a homogeneous group of low self-monitors where no differences in degree could be observed. These subjects are considered to be unscalable. However, two Rasch homogeneous latent classes were identified for the high self-monitors, where the main differences could be found with respect to "actor-type" items. Vansteelandt and Van Mechelen (2004) studied individual differences in situation-behavior profiles of anger. They derived three types of persons that could be characterized in terms of distinctive behavioral signatures. Another example shows the detection and specification of subtypes of depression, as opposed to classifying depression on either a continuum or on a typological level (Hong & Min, 2007). The structure of personality might be more complex than expected. A mixture IRT model may provide a better understanding of the data, without having to make a choice in describing personality in terms of types or traits.

The combination of kind and degree into one model may yield an impulse to the ongoing debate of situational specificity of personality and behavior. It was stated by Mischel (1968) that "Individuals show far less cross-situational consistency in their behavior than has been assumed by trait-state theories" (p. 177). Bem and Allen (1974) had a different view, and stated a number of reasons why intuition may be right, which tells us that subjects display cross-situational consistencies in their behavior. Empirical research has demonstrated that subjects exhibit differing degrees of consistency in their behavior across different situations. It is argued that some subjects are more consistent than others in their cross-situational behavior, which may influence the predictability of subjects (Bem & Allen, 1974; Berdie, 1961; Lanning, 1988; Reise & Waller, 1993). Thus, subjects may be scalable according to different models that take situational specificity of behavior into account.

A mixture IRT model may identify groups of subjects that respond to the items differently, depending on the situation. This focus is adopted in Chapter 2, where the

concept of self-disclosure is studied. It is expected that apart from an overall tendency to self-disclose that is reflected by the latent trait, groups of subjects differ in their pattern of self-disclosure with respect to the recipient of the disclosure (Omarzu, 2000; Steel, 1991). In this study, situational specificity concerns the person the subject is facing. Thus, subjects may be inconsistent in their behavior, reflecting their true nature rather than measurement error. From this perspective, qualitative differences described by the mixture IRT model may map situational specificity.

1.2 Qualitative Differences as Artifacts

Qualitative differences in response patterns can also be viewed as methodological artifacts. There are many variables unrelated to the measured construct that may affect items to function differently for different (groups of) subjects.

1.2.1 Differential Item Funtioning

An item is said to exhibit differential item functioning (DIF) when it has different response probabilities for different groups, after matching the groups with respect to their position on the latent trait (Angoff, 1993). Usually, methods for detection of DIF compare the functioning of an item across manifest groups (e.g. Camilli & Shepard, 1994; Holland & Wainer, 1993). The model assumes that these manifest variables represent homogeneous subgroups which are associated with the source of the DIF. However, several studies using mixture IRT models have shown that the manifest groups used for DIF detection are not necessarily identical to the groups of subjects for whom an item may be biased (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005).

Because group membership may be difficult to observe, Kelderman (1989) pointed to the inclusion of unobserved groups for detection of item bias. DIF detection methods using mixture IRT models focus on the differential functioning of an item across latent instead of manifest groups. Kelderman and Macready (1990) were the first to use the mixture IRT modeling framework to assess DIF, comparing DIF detection across manifest and latent groups. An advantage of a latent grouping variable is that it can be applied even when there is no manifest variable available, or when the manifest variable is not a valid indicator of the source of bias. However, if the manifest variable is a valid indicator

it is often measured with error, which can be modeled by including a latent class variable. Furthermore, a latent grouping variable allows for direct interpretation of possible causes of DIF, without being restricted to a specific manifest variable. For example, Cohen and Bolt (2005) showed that the mixture Rasch model could identify biased items in a mathematics placement test. The differential functioning of items was related to the type of mathematics problem.

De Ayala, Kim, Stapleton and Dayton (2002) compared several manifest DIF detection methods to assess DIF in data that were based on a mixture IRT model. It was shown that when the association between the source of bias and the manifest variable is reduced, the performance of the manifest DIF detection methods decreased rapidly. However, they did not study the performance of the mixture IRT modeling framework as a method for DIF detection.

The application of mixture IRT models as a method for identifying DIF items, is studied in Chapter 3. It is examined whether mixture IRT models, with a latent grouping variable, perform better in identifying DIF items compared to DIF detection methods using manifest variables only. A simulation study is conducted to compare the effectiveness of the latent and manifest DIF detection methods. The strength of association between the source of bias and the manifest variable is varied, as well as sample size.

1.2.2 Differential use of response scale

We have discussed qualitative differences in terms of artifacts that influence the functioning of a few items. However, qualitative differences can also be interpreted as measuring the same construct in different ways. In this case, the latent classes may be characterized by a differential use of the response scale, which may be substantively meaningful. Though unrelated to the measured construct, it is an important object of study as it has been shown to affect personality measurement. A number of studies have applied mixture IRT models to identify latent classes associated with differences in response behavior on self-report questionnaires. For example, Koeller (1994) applied the mixture Rasch model to identify guessing behavior among subjects. Furthermore, Zickar, Gibby, and Robie (2004) have shown to uncover faking subgroups by applying the mixture Rasch model. Moreover, they concluded that the traditional assumption that applicants fake and incumbents respond

honestly may be incorrect.

Differential use of the response scale may also manifest itself in differences in preference for specific categories (Jackson & Messick, 1958; Nunnally, 1967), like for the extreme ends of the response scale (Berg & Collier, 1953; Hamilton, 1968). Studies using the mixture Rasch model to analyze personality scales derived two latent classes (Austin, Deary, & Egan, 2006; Rost, Carstensen, & Von Davier, 1997). The latent classes could be interpreted in terms of subjects with a preference for the extreme ends of the response scale and a class of subjects preferring to use the middle of the scale. Comparable results were found in responses to a job satisfaction survey (Eid & Rauber, 2000). A mixture version of the nominal response model has been used to investigate differences in the use of response categories of multiple-choice items (Bolt, Cohen, & Wollack, 2001). In an English usage test, groups of subjects were distinguished that were drawn to different types of distractor responses.

Mixture IRT models have been used to study the differential meaning or preference for a "?" response category as well (Smit, Kelderman, & Van der Flier, 2003). Smit, Kelderman, and Van der Flier (2003) studied two personality scales with response categories "yes", "?", and "no". The parameters of the "?" category were not invariant over subjects, where subjects from one latent class tended to avoid this category. Similar results were found by Hernández, Drasgow, and González-Romá (2004). In Chapter 4, the study by Smit et al. (2003) is extended by analyzing a larger dataset, and allowing the identification of more than two latent classes. Response tendencies may result in a differential use of the response scale by different groups of subjects. Therefore, a mixture version of the nominal response model is used to analyze the data, where no a priori restrictions on the item parameters have to be made. A more parsimonious model is used for analysis as well.

Including manifest variables in a mixture IRT model may aid in the interpretation of latent classes. An example of a variable that may be associated with qualitative differences in response behavior is social desirability (Edwards, 1957; Paulhus, 1984; 1986). This concerns the deliberate response distortion in order to look good and make a favorable impression with others, as well as an unconscious, stable positive view of oneself. Cross-cultural research has pointed out the differences in response distortion across cultures. For example, Bachman and O'Malley (1984) reported on the black and white differences in

yea/nay-saying and extreme responding. For other examples see Hui and Triandis (1989), Grimm and Church (1999), and Johnson, Kulesa, Cho, and Shavitt (2005).

Van de Vijver and Phalet (2004) stressed the importance of the role of acculturation in assessment in multicultural groups, which is often overlooked (see also Van Hemert, Baerveldt, & Vermande, 2001). Although ethnic groups may be fairly heterogeneous due to differences in acculturation of subjects, ethnicity is incorporated in the model as a covariate to study its association with the latent class variable. Social desirability is added to the model as a covariate as well. In Chapter 4, the focus is on differences in response behavior across latent classes and interpretation in terms of response tendencies. In addition, latent classes are characterized by their association with the covariates.

1.3 Prediction

Because mixture IRT models describe the within as well as between-group differences, it can be argued that these models provide a more complete description of test behavior. Therefore, it can be expected that compared to a simple IRT model, mixture IRT models may be used to improve the accuracy of prediction. Recently, Eid and Zickar (2007) argued "In order to determine whether traits estimated using mixture distribution IRT are indeed more useful statistics, it is important to evaluate the predictive validity of those trait estimates compared to other types of trait estimates" (p. 268). To evaluate the predictive validity, predictions of external criteria by latent trait values estimated with a mixture IRT model can be compared to prediction by latent trait values estimated with a simple IRT model.

In Chapter 4, we evaluate the predictive validity of the latent trait estimates, by investigating the associations of these trait estimates with an external criterion measure. This external criterion is an assessment by a psychologist on the same construct as the questionnaire intends to measure. It is available for a part of the sample from the study described in Chapter 4. The possible improvement in prediction using a mixture IRT model compared to a simple IRT model is evaluated.

There is a general conception that each subject belongs to one, and only one, latent class (Goodman, 1974). Subjects are usually assigned to a latent class based on the highest probability given a subjects response pattern. Next, they obtain the latent trait

estimate corresponding to the latent class they have been assigned to. This means that the certainty of latent class assignment is not taken into account. For representation and characterization of latent classes this may be very convenient. Whether the procedure yields an optimal estimation of latent trait values and an optimal prediction of external criteria is another case.

Consider an example of a mixture IRT model with two latent classes. One subject has probabilities of .10 and .90 to belong to latent class one and two respectively. For a second subject, these probabilities are .45 and .55. This results in the assignment of both subjects to latent class two, but with a different degree of certainty of assignment. For both subjects, the item parameters of the second latent class can be used to estimate their latent trait value. Based on the individual latent class probabilities, it could be expected that this may be an accurate estimation of the latent trait value for the first subject, but a less accurate estimate for the second subject. A reason for the difference in accuracy of estimation is that the response pattern of the second subject may not be as well described by the IRT model of the second latent class as that of the first subject. Furthermore, it is ignored that the second subject also had a relatively high probability to be assigned to latent class one, whereas this may contain important information. Therefore, we study an alternative to assignment of latent trait estimates. The class-specific latent trait estimates are in that case weighted by their corresponding latent class probabilities. The weighting procedure is compared to the unweighted method in Chapter 5. Assignment may be necessary for description, but for prediction it may be better to use a weighted latent trait estimate as opposed to an assigned estimate.

1.4 Outline

In this thesis, several mixture IRT models are used to analyze the response behavior of subjects. The possible use of the mixture IRT modeling framework for various applications is investigated. Within and between-group differences are modeled simultaneously. Quantitative individual differences are reflected by the latent trait, which describes differences in degree to which subjects have a certain attribute. Qualitative differences described by the latent class variable are discussed from different perspectives. They may relate to differences in kind, associated with situational specificity of behavior as described

in Chapter 2. Also, response behavior unrelated to the measured construct is studied, where qualitative differences are considered to be methodological artifacts. In Chapter 3, the mixture IRT model will be studied as a means for identifying items exhibiting DIF. Furthermore, it can be investigated whether certain tendencies influence the response behavior of (some of) the subjects, and what kind of response tendencies these are. This is the focus of Chapter 4, where there will also be a first look at the accuracy of prediction using mixture IRT models. It is examined whether mixture IRT models provide means to improve prediction in comparison with a simple IRT model. Different scoring rules are inspected in Chapter 5, to study prediction using weighted versus assigned latent trait estimates.

Chapter 2

Latent-Trait Latent-Class Analysis of Self-disclosure in the Work Environment

Based on the literature about self-disclosure, it was hypothesized that different groups of subjects differ in their pattern of self-disclosure with respect to different areas of social interaction. An extended latent-trait latent-class model was proposed to describe these general patterns of self-disclosure. The model was used to analyze the data of 1113 subjects, tested on extraversion and with respect to their degree of self-disclosure towards different categories of people in the work environment. A model with one latent trait and a latent-class variable with three categories was identified. Subjects belonging to the different latent classes differ in their general tendency to self-disclose, in their choice to whom they will show self-disclosure and in the degree to which they are selective in their self-disclosure. The collateral variable extraversion was associated with both latent variables. The association of extraversion with selectivity in self-disclosure was not significant.

The concept of self-disclosure has had a long history. Jourard and Lasakow (1958) refer to self-disclosure as "the process of making the self known to other persons". Cozby (1973) defines self-disclosure as "any information about himself which person A communicates verbally to person B". According to Cozby (1973) and Omarzu (2000) self-disclosure consists of three basic dimensions. The first is the breadth or amount of information

This chapter is a minor revised reprint of: Maij-de Meij, A.M., Kelderman, H., & Van der Flier H. (2005) Latent-trait latent-class analysis of self-disclosure in the work environment. *Multivariate Behavioral Research, 40, 435-459.*

disclosed, referring to the number of topics covered by the disclosure. The depth or intimacy of the information disclosed is the second dimension. The third is the duration or time spent describing each item of information.

Jourard was one of the first researchers who operationalized self-disclosure. In collaboration with Lasakow he developed the Self-Disclosure Questionnaire (SDQ, Jourard & Lasakow, 1958). Jourard intended to use the scale to identify the larger social patterns of disclosure content, as well as individual, trait-like differences in self-disclosure tendencies. He conjectured that differences in self-disclosure are determined above all by stable personality differences (Jourard, 1971).

Self-disclosure has been studied not only as a personality construct but also as a behavioral process occurring during interaction with others. Aspects of this process, such as reciprocity and social exchange, have been studied extensively (Cozby, 1973; Morton, 1978; Rubin, 1975). These aspects have to do with the development of social relationships (Cozby, 1973). Altman and Taylor (1973) developed the social penetration model. This model describes how social relationships between strangers develop from casual acquaintanceships to close personal friendships. Other studies focused on who elicits self-disclosure from others (Colvin & Longueuil, 2001; Miller, Berg, & Archer, 1983). In this paper, self-disclosure is seen from the perspective of a personality construct. It is studied whether there are qualitative individual differences in self-disclosure patterns.

Self-disclosure is often found to be related to extraversion. Several studies show the degree of self-disclosure to be correlated to personality measures. See Cozby (1973) for an overview. Extraversion can be defined as the degree to which one's energy, attention and orientation is directed outwards. An extravert person is someone who is not shy and prefers to spend time with other people rather than alone, and who has an active involvement with the environment. Introvert people on the other hand, have more negative expectations about social interactions, which can lead to social avoidance. They tend to be on their own, and to withdraw into themselves (Carver & Scheier, 1995; Morris, 1979).

2.1 Selectivity in Self-disclosure

Omarzu (2000) developed the Disclosure Decision Model for the processes that determine the specific dimensions of individuals' disclosure. Whether any self-disclosure will be

made in a given situation depends on the presence of social goals. These goals can be social rewards that one can achieve through self-disclosure. Which goal is important to someone, depends on the individual. Also, situational cues must highlight the salience of the particular social reward. Next, it is decided whether self-disclosure is an appropriate strategy to exercise and to whom one will disclose, otherwise the self-disclosure strategy has no satisfactory goal utility (Miller & Read, 1987; Omarzu, 2000).

The final decision has to be made regarding precisely what to disclose. The model assumes that people evaluate the utility of disclosure rewards as well as the risks of self-disclosure. These risks include, among others, social rejection, betrayal, and causing discomfort to the listener (Omarzu, 2000). Omarzu (2000) hypothesized that, "As subjective risk increases, the depth of disclosure will decrease. ... Even when the subjective utility of the goal is high, perceived risk should decrease the emotional intensity of disclosures" (p. 180).

This is in agreement with Steel's (1991) finding that interpersonal trust and self-disclosure are positively related. When people trust others and do not feel they can be hurt easily, then they will show self-disclosure. These people do not see much risk in self-disclosure because they are less suspicious and prejudiced than people who do not trust others that easily. Therefore, their self-disclosure is often deep and more intimate (Omarzu, 2000). Because more extravert people have less suspicion towards others and feel more interpersonal trust, it can be expected that they have a greater tendency to disclose themselves to others. Furthermore, because they are about equally open to different categories of people, their trust depends on their personality rather than on the person in front of them. Therefore, it is expected that they will be less selective in the person to whom they will self-disclose to. On the other hand, introvert people see much risk in self-disclosure; they have negative expectations about social interaction. They will be less trusting and will not show much self-disclosure. Also, they will be more selective in their choice to whom they will show self-disclosure.

The Disclosure Decision Model leads to the introduction of the concept of selectivity in self-disclosure. This would be an extension of the literature on self-disclosure by drawing attention to situational specificity. In this study situational specificity concerns the person the subject is facing. If the subject doesn't know a person, possible prejudice and suspicion will be based on rough social categories (Vonk, 1999). In the work environment these

categories are primarily: employees, colleagues, superiors, and customers. Because the prejudices are stable over time, selectivity in self-disclosure with respect to these categories of people will also be stable over time.

To determine to what extent people are selective in their self-disclosure, one has to look at the differences in their self-disclosure towards different (groups of) people to which they are exposed. When differences in self-disclosure with respect to the different categories of people are small, it means that the person is equally open and will take an equally vulnerable position towards the different categories of people. To these subjects it doesn't matter who they are facing, they are not selective in their self-disclosure. When the differences in self-disclosure are large, it indicates that it matters to the person who (s)he is facing. This person is not equally open to everybody, but is more selective in the choice to whom (s)he will show self-disclosure.

In summary, it can be expected that there is an overall tendency to self-disclose that is reflected in different areas of social interaction. It is hypothesized, however, that different groups of subjects differ in their pattern of self-disclosure with respect to the different areas of social interaction. Subjects who respond differently to the different (groups of) people, are selective in their self-disclosure. So, different response patterns of self-disclosure may reflect differences in selectivity in self-disclosure. It is expected that there are qualitative individual differences with respect to self-disclosure reflecting differences in selectivity in self-disclosure (Hypothesis 1). In this paper, mixture measurement models are specified to test whether sub-populations can be distinguished, that have qualitative different response patterns.

Hypothesis 2, concerns the relationship between extraversion and self-disclosure. It is expected that people who are extravert, will show more self-disclosure than introvert people (a). Concerning selectivity in self-disclosure, it is expected that in comparison to introvert people, extravert people are less selective in their self-disclosure (b). In the next section a model is developed that is suitable to determine subgroups with different patterns of self-disclosure but also allows self-disclosure responses to be associated within each subgroup.

2.2 General modeling framework

The general modeling framework is an extensions of mixture measurement models introduced by Rost (1990; 1991), Kelderman and Macready (1990), Mislevy and Verhelst

(1990), and Heinen (1996). First, associations between the responses are modeled with Bock's (1972) nominal response model. The model is written as a latent-class association model, where the continuous latent trait is made discrete (Heinen, 1996). Bock's nominal response model assumes that the responses of all subjects in the population of interest are governed by the same measurement model. However, since the primary interest is in detecting subgroups of subjects that show different patterns of self-disclosure with respect to the recipient of the disclosure, we propose a mixture model. The mixture components correspond to the subgroups, and the patterns of difficulty parameters in Bock's model correspond to patterns of self-disclosure. The model is parameterized such that the latent trait takes care of the common variation of responses within each mixture component. In the sequel, the mixture components are called latent classes and the quantitative latent variable of Bock's nominal response model is the 'latent trait'. Finally, to study whether extraversion is associated with self-disclosure, the model is extended with a collateral variable that is exogenous with respect to the latent variables. The final model is a discrete recursive graphical model (Cox & Wermuth, 1996; Lauritzen, 1996) containing latent variables. Figure 2.1 depicts the graphical model for the self-disclosure data. The arrows denote a regression, the squares represent the manifest variables, while the ellipses represent the latent variables. The latent class variable is allowed to influence the difficulty parameter for the relation between the self-disclosure responses and the latent trait.

First, the relations between the latent variables and the self-disclosure responses are modeled. Let θ denote a latent trait value describing the degree to which a subject possesses the attribute of interest. Furthermore, let X_j denote a subject's response to

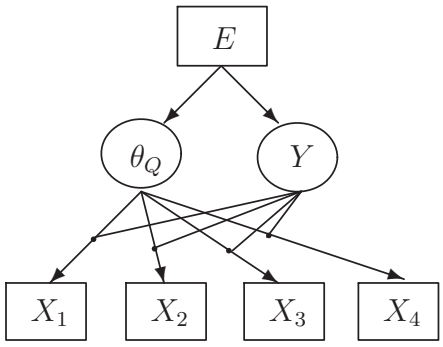

Figure 2.1: Discrete recursive graphical model with variables extraversion, E, latent trait, θ_Q, mixture component, Y, self-disclosure towards employees, X_1, colleagues, X_2, superiors, X_3, and customers, X_4.

item $j \in \{1, \ldots, J\}$ taking values $x_j = k \in \{1, \ldots, K_j\}$. Bock's (1972) nominal response model (NRM) describes the probability of x_j given θ as

$$P(X_j = k|\theta) = \frac{\exp(c_{jk} + a_{jk}\theta)}{\sum_{h=1}^{K_j} \exp(c_{jh} + a_{jh}\theta)}, \tag{2.1}$$

where the intercept c_{jk} and slope a_{jk} are restricted to sum to zero over the responses. In item response theory models, the slope parameter is called the 'discrimination' parameter, and the category intercept is called the 'difficulty' parameter. The denominator is a constant of proportionality which ensures that the probabilities sum to one over the item responses. The NRM can be re-formulated as a row-column association model (Goodman, 1979; Heinen, 1996).

First write Equation 2.1 as a loglinear model, that is,

$$\log P(X_j = k|\theta) = d + c_{jk} + a_{jk}\theta, \tag{2.2}$$

where d is $-\log$ the proportionality constant. The slope parameter is re-parameterized as,

$$a_{jk} = \tau_{jk}\xi_j^{J\Theta} \tag{2.3}$$

such that τ_{jk} satisfies both

$$\sum_{k=1}^{K_j} \tau_{jk} = 0$$

and

$$\sum_{k=1}^{K_j} \tau_{jk}^2 = 1. \tag{2.4}$$

From Equation 2.3 and 2.4 one has

$$\xi_j^{J\Theta} = \{(1\xi_j^{J\Theta})^2\}^{\frac{1}{2}} = \{\sum_{k=1}^{K_j} \tau_{jk}^2 (\xi_j^{J\Theta})^2\}^{\frac{1}{2}} = \{\sum_{k=1}^{K_j} \tau_{jk}\xi_j^{J\Theta})^2\}^{\frac{1}{2}} = (\sum_{k=1}^{K_j} a_{jk}^2)^{\frac{1}{2}}, \tag{2.5}$$

where the second equation follows from Equation 2.4 and the fourth equation follows from 2.3. The parameter $\xi_j^{J\Theta}$ describes the overall association between the response to item j and the latent trait, while the parameter τ_{jk} represents a category score for the response x_j. If it is assumed that subjects are randomly selected from a population with distribution $f(\Theta)$, one has from Equation 2.2 and 2.3

$$\begin{aligned} \log P(X_j = k, \theta) &= \log\{P(X_j = k|\theta)f(\theta)\} \\ &= d + \log f(\theta) + c_{jk} + \tau_{jk}\xi_j^{J\Theta}\theta. \end{aligned}$$

The distribution $f(\Theta)$ can be approximated to any degree of accuracy by a discrete distribution $g(\theta_Q)$, where Q is a nominal variable with categories $q = (1, \ldots, r)$, and θ_q is a fixed metric score assigned to category q. For simplicity, it is assumed that the scores θ_q have equal distances (Heinen, 1996). The joint loglinear model for the discrete latent trait variable and the item responses now becomes

$$\begin{aligned} \log P(X_j = k, \theta_Q) &= \log\{P(X_j = k|\theta_Q)g(\theta_Q)\} \\ &= \lambda + \lambda_q^Q + \lambda_{jk}^{JK} + \mu_{jk}\phi_j^{JQ}\theta_q, \end{aligned} \tag{2.6}$$

For simplicity, from here on the superscripts will be omitted, except where needed. The parameter λ_{jk} denotes the item difficulty parameter. The superscripts of the association parameter, ϕ_j^{JQ}, are markers, indicating the variables for which the association is defined. To obtain estimable category scores, the item category scores μ_{jk} are restricted to satisfy

$$\sum_{k=1}^{K_j} \mu_{jk} = 0, \qquad \text{and} \qquad \sum_{k=1}^{K_j} \mu_{jk}^2 = 1.$$

The category scores scale the categories of the items and provide information about the distances between the response categories. The odds ratio of the distances between the response categories gives information about the intervals between the categories. By estimating the item category scores, no assumptions have to be made with respect to the ordering of the categories of the manifest variables. If they are properly ordered, this will be reflected in their estimated values (Clogg, 1982; Goodman, 1979). Without loss of generality, the latent trait scores are fixed in advance and chosen to satisfy

$$\sum_{q=1}^{r} \theta_q = 0, \qquad \text{and} \qquad \sum_{q=1}^{r} \theta_q^2 = 1.$$

Finally, if the main effect of the latent trait is parameterized to sum to zero over the index, then it relates to the distribution $g(\theta_Q = \theta_q)$ by

$$\lambda_q = \log g(\theta_Q) - r^{-1} \sum_{q=1}^{r} \log g(\theta_Q).$$

Model Equation 2.6 is a latent trait model where the observed self-disclosure responses are related to the discrete latent trait via a row-association model (Goodman, 1979; see also Anderson & Vermunt, 2000). The model is easily extended to a mixture latent-trait

latent-class model by adding a latent class variable Y, where the values m represent the mixture components,

$$\begin{aligned} \log P(X_j = k, \theta_Q, Y = m) &= \log\{P(X_j = k|\theta_Q, Y = m)g(\theta_Q)P(Y = m)\} \\ &= \lambda + \lambda_q + \lambda_m + \lambda_{jk} + \lambda_{jkm} + \mu_{jk}\phi_j^{JQ}\theta_q, \end{aligned} \tag{2.7}$$

with additional identifying restrictions, $\sum_m \lambda_m = 0$ and $\sum_m \lambda_{jkm} = 0$. The general mean is denoted by λ, whereas the main effects of the discrete latent trait, the latent class and the self-disclosure responses x_j are represented by λ_q,λ_m, and λ_{jk} respectively. The parameter λ_{jkm} describes the interaction of self-disclosure score x_j and latent class membership. Note that the sum λ_{jk}+λ_{jkm} is equal to the class-specific difficulty parameter of response x_j in the nominal response model of Equation 2.1. The model term λ_{jkm} describes the between class differences, whereas the term $\mu_{jk}\phi_j^{JQ}\theta_q$ describes the individual differences in self-disclosure responses within each class. The association between the response to item j and the latent trait is described by the parameter ϕ_j^{JQ}.

Model Equation 2.7 describes the relations between self-disclosure responses and the latent variables. That is the lower half of Figure 2.1. In the upper half of this figure, the collateral variable extraversion is related to the latent variables. Both latent class membership and the distribution of the latent trait are specified conditional on the collateral variable. Extraversion is related to the item responses via the latent variables.

A row-column association model is formulated to model the relation between the latent class variable and extraversion, while the relation with the discrete latent trait is specified by a row-association model (Goodman, 1979). Adding these associations, as well as a main effect of extraversion, the joint loglinear model for all latent and manifest variables becomes,

$$\begin{aligned} &\log P(X_j = k, \theta_Q, Y = m, E = e) \\ &\quad = \log\{P(X_j = k|\theta_Q, Y = m)g(\theta_Q|E = e)P(Y = m|E = e)P(E = e)\} \\ &\quad = \lambda + \lambda_q + \lambda_m + \lambda_e + \lambda_{jk} + \lambda_{jkm} + \mu_e\phi^{EQ}\theta_q + \nu_e\phi^{EY}\mu_m + \mu_{jk}\phi_j^{JQ}\theta_q, \end{aligned} \tag{2.8}$$

with additional identifying restrictions $\sum_e \lambda_e = 0$, $\sum_e \mu_e = \sum_e \nu_e = \sum_m \mu_m = 0$ and $\sum_e \mu_e^2 = \sum_e \nu_e^2 = \sum_m \mu_m^2 = 1$. The category score of extraversion in the relation to the latent trait is denoted by μ_e, whereas the category scores in the relation between extraversion and the latent classes are represented by ν_e and μ_m, for extraversion and the

latent classes respectively. Furthermore, the association parameters ϕ^{EQ} and ϕ^{EY} describe the association between the latent trait and extraversion and between the latent class variable and extraversion respectively. The category scores and association parameters will be estimated in the model. Simulation research shows that the standard errors of the parameter estimates as well as latent class assignment can benefit substantially from incorporating collateral variables (Smit, Kelderman, & Van der Flier, 1999; 2000).

Finally, assuming conditional independence of the self-disclosure responses $\boldsymbol{X} = (X_1, \ldots, X_j)$ given the latent trait and latent class membership, one obtains from Equation 2.8,

$$\begin{aligned} &\log P(\boldsymbol{X} = \boldsymbol{x}, \theta_Q, Y = m, E = e) \qquad (2.9) \\ &= \lambda + \lambda_q + \lambda_m + \lambda_e + \sum_{k=1}^{K_j} \lambda_{jk} + \sum_{k=1}^{K_j} \lambda_{jkm} + \mu_e \phi^{EQ} \theta_q + \nu_e \phi^{EM} \mu_m + \sum_{k=1}^{K_j} \mu_{jk} \phi_j^{JQ} \theta_q, \end{aligned}$$

where $\boldsymbol{x} = (x_1, \ldots, x_j)$ are the values taken by $\boldsymbol{X}$. The sum $\lambda_{jk} + \lambda_{jkm}$ corresponds to the item difficulty parameter of the nominal response model as described in Equation 2.1, which may vary over latent classes. The item discrimination parameter of the model corresponds to $\mu_{jk}\phi_j^{JQ}$, where the association parameter ϕ_j^{JQ} may vary over the items, and the category scores μ_{jk} over the items and their categories.

If $n_{\boldsymbol{x}e}$ denotes the observed frequency of the manifest responses $\{x_1, \ldots, x_j, e\}$, the log-likelihood of the model given by Equation 2.9 can be written as

$$\begin{aligned} L &= \log \prod_e \prod_{\boldsymbol{x}} \{P(\boldsymbol{X} = \boldsymbol{x}, E = e)^{n_{\boldsymbol{x}e}}\} \\ &= \sum_e \sum_{\boldsymbol{x}} n_{\boldsymbol{x}e} \log \sum_{\theta_q} \sum_m P(\boldsymbol{X} = \boldsymbol{x}, \theta_Q, Y = m, E = e). \end{aligned}$$

If ζ denotes the vector of independent parameters in Equation 2.9, the maximum likelihood equations are obtained by solving

$$\frac{\partial L}{\partial \zeta} = 0.$$

The maximum likelihood estimates of the model parameters are computed by means of the EM-algorithm (Dempster, Laird, & Rubin, 1977). In the E(xpectation)-step, the probabilities for the complete data matrix are estimated given the observed data and the parameter estimates. This is followed by the M(aximization)-step, where the log-likelihood for the complete data matrix is maximized to obtain new estimates for the

model parameters. The algorithm is repeatedly applied until some convergence criterion is met.

If a model is a special case of another model and not on the border of the parameter space of that model, the difference in L^2-statistics with the corresponding degrees of freedom, equal to the difference in degrees of freedom of both models, can be used to compare the relative fit of the two models, and to determine which of the two models has the best fit to the data (Goodman, 1979). In the case of comparing models with different numbers of latent classes, the information criteria should be used. This is done by comparing the well known AIC-(L^2) statistics of two models. The best solution, defined the lowest value of the AIC-statistics, will be chosen. Since the models may have local maxima, they are analyzed several times with different sets of random starting values. All analysis are performed with the program ℓEM (Vermunt, 1997).

2.3 Method

2.3.1 Subjects

A total of 1113 subjects, 811 men and 302 women, were tested between October 1999 and February 2002 in connection with a personnel selection or a personal development program at a Dutch consultancy firm, dealing with organizational development and recruitment & personnel selection. The educational level attained by the subjects was that of high school and/or higher education. They were employed or applying for middle to upper level positions in the service providing industry; these are operational and commercial functions.

2.3.2 Instrumentation

Self-disclosure is measured by a Dutch computer-administered questionnaire (Blom, 1992) measuring self-disclosure with respect to four different categories of people: employee (X_1), colleague (X_2), superior (X_3), and customer (X_4). There are forty items, ten for each type of self-disclosure. The four sub-scales are parallel versions; they differ in who the other person is. Each item states a situation and gives two options describing different ways to react to the particular situation. One option describes the tendency to share feelings and opinions with the other person. The other option indicates the reverse. The

subject has to indicate with which of the two options he/she agrees most on a six point scale. An example of an item from the aspect self-disclosure towards an employee is:

> Due to repeated discussions about an important subject the relationship between you and one of your employees has deteriorated.
>
> I During our contacts I avoid the subject as much as possible in order not to disturb our relationship even more.
>
> II I mention the effect this conflict has on our relationship and suggest that we talk this out straight away.

Option I does not indicate self-disclosure with respect to the conversation partner, while option II does. The reliabilities (values of Cronbach's alpha) of the sub-scales vary between .63 and .72. The attention is focussed on differences in the score pattern over the self-disclosure variables, not on differences in response patterns within each self-disclosure variable. Therefore, for each subject the four scale scores of the self-disclosure variables are used, describing self-disclosure towards the different categories of people. The scale scores had a range from 1 to 10, with the scores 1 and 10 occurring only in a small part of the sample. To prevent estimation problems, the scores 1 and 2 were combined. The same was done for the scores 9 and 10, to obtain 8 categories with a sufficiently large number of subjects in each category. The scale from 1 to 8 corresponds to not having the characteristic at all and having this characteristic to a high degree. The correlations among the four sub-scales are between .51 and .70.

The other measure of interest is extraversion (E). Extraversion is measured by nine items selected from the extraversion scale of the Dutch adaptation of the NEO-PI questionnaire (Hoekstra, Ormel, & De Fruyt, 1995). Each of the nine items consists of a statement for which one has to indicate the degree to which one agrees with it. There are five options: agree completely, agree, neutral, disagree and disagree completely. An example of an item of the aspect extraversion is:

> I love having people around me.

The value of Cronbach's alpha of this scale is .80. Here, the scale score of extraversion is used as well, which is in accordance with the scales of the self-disclosure measures.

2.3.3 Analyses

First the model as formulated in Equation 2.9 will be analyzed with different numbers of latent classes. Furthermore, when the number of latent classes that gives the best fit to the data is determined, it will be analyzed whether the associations between the responses, as modeled with the latent trait, are class specific, that is whether $\phi_j^{JQ} \neq \phi_{jm}^{JQ}$. Finally, it will be examined whether the assumption of conditional independence of the latent trait and the latent class variable holds.

The first hypothesis implies that there should be more than one latent class. When the model with the best fit to the data includes a latent class variable with more than one latent class, it is demonstrated that subpopulations can be distinguished that have qualitative different response patterns. To determine whether the differences in response patterns with respect to the self-disclosure variables reflect differences in selectivity in self-disclosure, the characteristics of the latent classes have to be examined.

First, it will be examined how the self-disclosure responses are associated with the latent trait, by looking at the association parameters, ϕ_j^{JQ}, and the category scores μ_{jk} of the self-disclosure responses x_j. To characterize the latent classes, first the λ_{jkm} parameters will be examined. These parameters show the tendency to self-disclose towards the different categories of people for each latent class. A positive value indicates a higher tendency to respond in the specific response category, compared to the other response categories of the same item. Next, the expected category scores of the self-disclosure variables with respect to the different categories of people, given someone's score on the latent trait and latent class membership, will be computed by,

$$E(\mu_j|\theta_Q, Y = m) = \sum_{k=1}^{K_j} \mu_{jk} P(X_j = k|\theta_Q, Y = m), \tag{2.10}$$

where $P(X_j = k|\theta_Q, Y = m)$ is computed from the estimated model Equation 2.9 using elementary probability theory. Equation 2.10 describes the dependence of self-disclosure response x_j on the latent trait in each latent class.

To describe the patterns of expected category scores of the self-disclosure variables, given latent class membership,

$$E(\mu_j|Y = m) = \sum_{k=1}^{K_j} \mu_{jk} P(X_j = k|Y = m), \tag{2.11}$$

is computed. Using Equation 2.11, each latent class can be interpreted in terms of the general degree of self-disclosure relative to the other latent classes, as well as by the patterns across the four self-disclosure variables. To compute the variance over these expected category score over self-disclosure responses, let Z_j be a statistic computed for self-disclosure component j, and $\boldsymbol{Z} = (Z_1, \ldots, Z_j)$. Let $Var_J(\boldsymbol{Z})$ denote the variance over the set J of the elements of $\boldsymbol{Z}$, thus,

$$Var_J(\boldsymbol{Z}) = \frac{\sum_{j=1}^{J}(Z_j - \overline{\boldsymbol{Z}}_J)^2}{j-1}, \tag{2.12}$$

where

$$\overline{\boldsymbol{Z}}_J = \frac{\sum_{j=1}^{J} Z_j}{j},$$

is the mean taken over J. By substituting $\boldsymbol{Z}$ in Equation 2.12 for each latent class, with a vector of four expected category scores of Equation 2.11, the variance of the expected category scores given latent class membership can be computed.

Hypothesis 1 stated that the different latent classes could be characterized by differences in the degree of selectivity in their self-disclosure. The degree of selectivity in self-disclosure depends on the subject's variability of self-disclosure responses towards the different categories of people. Therefore, the mean intra-person variance of each latent class will be computed, that is, the mean variance of the self-disclosure response patterns of the subjects belonging to each of the latent classes. If $\boldsymbol{\mu}_j$ denotes a vector of four self-disclosure category scores $(\mu_{1k}, \ldots, \mu_{jk})$, then the mean intra-person variance for each latent class is,

$$E[Var_J(\boldsymbol{\mu}_j)|Y = m] = \sum_j Var_J(\boldsymbol{\mu}_j) P(\boldsymbol{X} = \boldsymbol{x}|Y = m). \tag{2.13}$$

The latent class with the highest mean intra-person variance, consists of subjects who are the most selective in their self-disclosure towards the different categories of people, compared to the other latent classes. It is expected that subjects who show a low degree of self-disclosure, will be the most selective in their self-disclosure.

The second hypothesis implies that the latent class consisting of subjects who show a relatively high degree of self-disclosure, will also attain a relatively high score on extraversion. It will be examined whether the latent class showing the highest tendency to self-disclose, as shown by λ_{jkm}, will show the highest expected category score on

extraversion, and visa versa. Furthermore, it is expected that the category scores of extraversion, μ_e, in the relation with the latent trait are non-decreasing. Spearman's correlation coefficient will be computed for the relationship between the intra-person variance and extraversion, to examine whether subjects with low variation in their response patterns are more extravert.

2.4 Results

The results of goodness-of-fit testing of latent-trait latent-class model, see Equation 2.9, with different numbers of latent classes are given in Table 2.1. The overall likelihood-ratio chi-squared statistics (L^2) and their degrees of freedom (df) are given, as well as the corresponding AIC-statistics (AIC). The differences in the fit-statistics of the subsequent the models are given as well.

Table 2.1: *Fit Statistics of the Latent Class Analyses with a Discrete Latent Trait, a Latent Class Variable and Five Manifest Variables, as well as the Differences Between the Models*

Model	L^2	AIC	df	ΔL^2	ΔAIC	Δdf
1 - Model with Y = 1 classes	4739.722	-60642.277	32691			
2 - Model with Y = 2 classes	4658.809	-60663.191	32661	80.913	20.914	30
3 - Model with Y = 3 classes	4546.593	-60715.407	32631	112.216	52.216	30
4 - Model with Y = 4 classes	4515.954	-60686.046	32601	30.639	-29.361	30

On examining which number of latent classes gives the best explanation of the structure of the data, the AIC-statistics were compared. These results suggest that the model with three latent classes has the best fit to the data, and proves that a model with only the latent trait, one latent class, is not sufficient to explain the structure of the data. It is demonstrated that subpopulations can be distinguished that have qualitative different response patterns, which is a first support of the Hypothesis 1.

Furthermore, it is examined whether the association parameter of Model 3 should be class-specific, that is, whether $\phi_j^{JQ} \neq \phi_{jm}^{JQ}$. Finally, Model 3 with the addition of an association between the latent trait and the latent class variable is analyzed, to test the conditional independence of the two latent variables. Both models showed a solution on the boundary of the parameter space, indicated by a large number of zero estimated

frequencies. As a consequence, the asymptotic distribution of the L^2-statistics statistics is no longer a chi-square distribution, and therefore, the fit-statistics of these models cannot be trusted. Consequently, Model 3 was chosen as the best fitting model, where the latent trait and latent class variable can be considered to be independent.

2.4.1 Parameter Estimates

The probability of belonging to latent class one, .25, is the smallest. The probability of belonging to latent class two is .36, while for latent class three this is .39.

In Table 2.2 the results of the analysis of the associations between the latent trait and each of the self-disclosure variables are shown. The association parameters correspond

Table 2.2: *Category Scores, μ_{jk}, and Association Parameters, ϕ_j^{JQ}, of the Association Between the Latent Trait Variable and the Self-disclosure Variables*

Response Category	X_1	X_2	X_3	X_4
1	-0.4625	-0.6399	-0.5648	-0.4189
2	-0.3071	-0.2193	-0.3350	-0.2759
3	-0.2841	-0.1218	-0.2312	-0.2515
4	-0.1460	-0.0263	-0.0767	-0.1945
5	-0.0393	-0.0175	0.0313	-0.0331
6	0.2504	0.1033	0.2780	0.1448
7	0.3584	0.2471	0.3315	0.2956
8	0.6301	0.6744	0.5668	0.7334
Association Parameter	217.7704	620.4701	122.7333	141.8184

to the ϕ_j^{JQ} parameter of Equation 2.9, for response x_j to item j which in this study is one of the four self-disclosure variables X_j. The item category scores correspond to the μ_{jk} parameters of the same equation. It is expected that the ordering of the response categories, as described by the estimated item category scores, is non-decreasing.

The association parameters suggest that self-disclosure towards colleagues, X_2, relates stronger to the latent trait than the other self-disclosure variables. The item category scores are non-decreasing as one responds in a higher response category. The results of Table 2.2 are depicted in Figure 2.2. This is accomplished by taking the product of the category scores and the association parameters.

Table 2.3: *Parameter Estimates of λ_{jkm}, Describing the Relation Between the Latent Class Variable and the Self-disclosure Variables*

Response Category	Class											
	1				2				3			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	-1.85	0.51	-0.22	-0.57	1.85	5.06	2.44	1.03	NE*	-5.57	-2.22	-0.46
2	-1.46	-1.41	-0.93	-0.51	2.42	2.53	2.04	1.09	-0.96	-1.12	-1.11	-0.58
3	NE*	-1.70	-1.19	-0.85	0.34	0.84	1.09	1.00	-0.34	0.87	0.10	-0.16
4	-1.47	-1.77	-1.62	-1.03	0.58	0.19	0.54	0.23	0.90	1.58	1.08	0.79
5	-0.29	0.20	0.15	-0.51	-1.07	-2.67	-1.34	-0.34	1.36	2.47	1.19	0.85
6	1.40	0.71	0.81	0.48	-2.37	-2.47	-1.84	-1.40	0.97	1.77	1.03	0.92
7	3.08	2.92	0.24	1.65	-3.08	-3.20	NE*	-2.31	NE*	0.28	-0.24	0.65
8	2.58	0.00	2.51	1.68	NE*	NE*	-2.44	NE*	-2.58	NE*	-0.07	-1.68

* NE = not estimable.

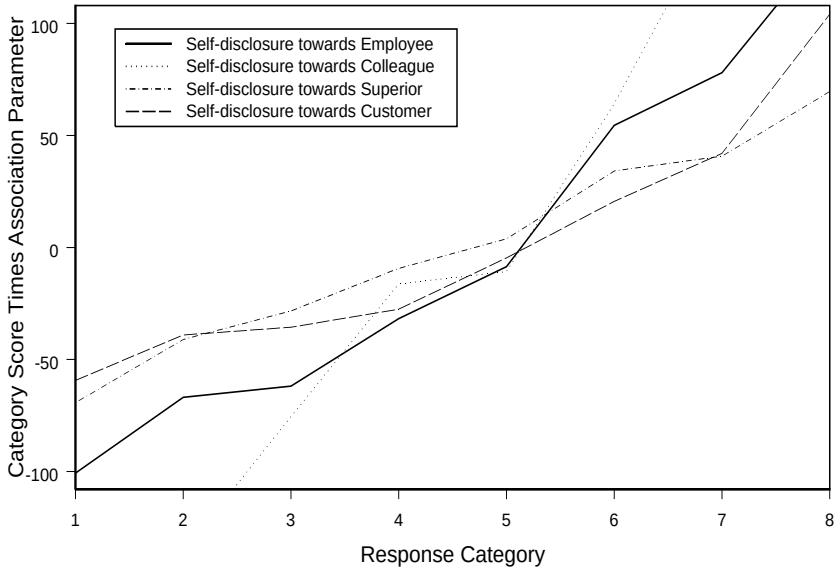

Figure 2.2: Category scores times the association parameter.

It is assumed that measured self-disclosure has an ordinal association with the latent trait variable, which describes latent self-disclosure. It is seen in Table 2.2, as well as in Figure 2.2, that the relation between measured self-disclosure and the latent trait is non-decreasing. The higher one scores on the latent trait variable, the higher one also scores on the self-disclosure variables.

The term λ_{jkm} in Equation 2.9 describes the tendency to respond in a certain response category, given latent class membership. Table 2.3 shows the corresponding parameter estimates. The NE entries in Table 2.3 are not estimated; there is insufficient information in the data to obtain adequate estimates under the model. It can be seen in Table 2.3, that latent class one is characterized by relatively high scores on self-disclosure towards the different categories of people. Subjects belonging to this latent class have a high tendency to show self-disclosure. The second latent class indicates the opposite. It consists of subjects with relatively low scores on self-disclosure. Latent class three shows medium scores (between 4 and 6) on the self-disclosure variables.

The next three graphs, in Figure 2.3, describe the expected category score of the self-disclosure variables with respect to the different categories of people, given someone's score on the latent trait variable and given the class someone belongs to, as computed by Equation 2.10. The graphs demonstrate that all latent classes indicate that the expected

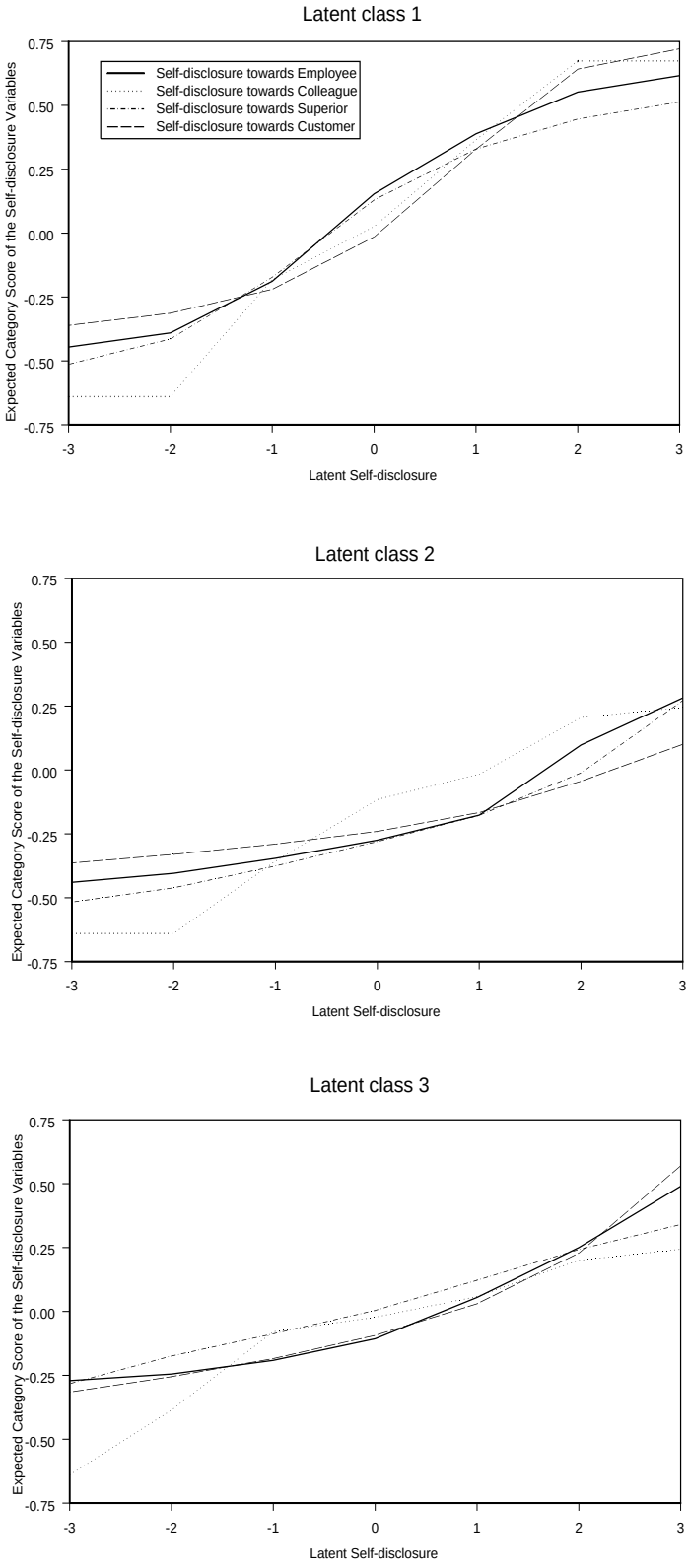

Figure 2.3: A graph for each latent class, describing the expected category score on measured self-disclosure with respect to the different categories of people as a function of the latent trait variable.

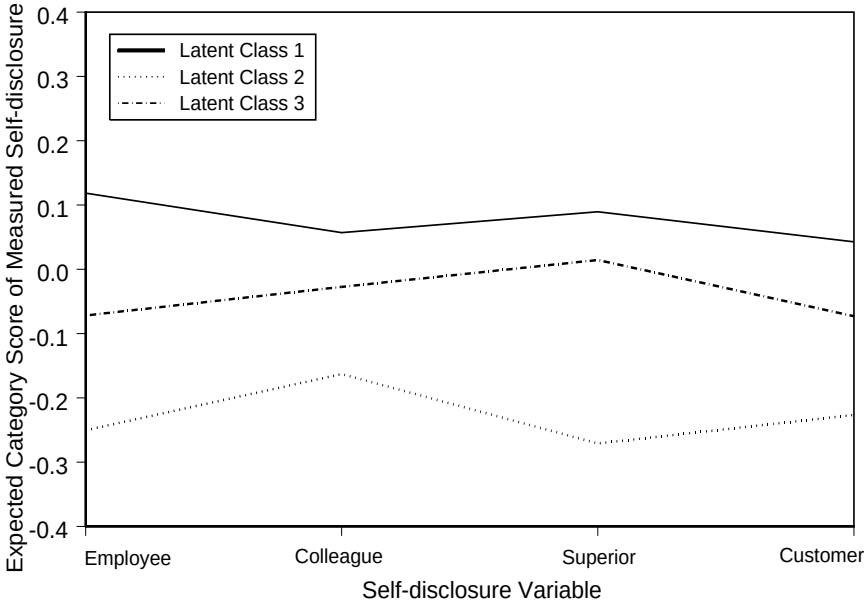

Figure 2.4: Expected category score on measured self-disclosure for each combination of a latent class and a self-disclosure variable.

category scores of the self-disclosure variables are increasing as one scores higher on the latent trait. The results also indicate that subjects belonging to the second latent class have relatively low scores on the self-disclosure variables with respect to the different categories of people. The subjects are relatively closed, they do not show much self-disclosure to others. Subjects belonging to latent class three are not likely to attain extreme scores, except for self-disclosure towards colleagues at the lower end of latent self-disclosure. Latent class one consists of subjects who may attain high scores. These results are in agreement with those of Table 2.3.

Figure 2.4 depicts for each self-disclosure variable the expected category score given latent class membership. Each line indicates a different latent class. The expected category scores are computed as shown by Equation 2.11. Figure 2.4 shows that the latent classes differ in their general level of measured self-disclosure and that the latent classes each have different response patterns with respect to self-disclosure towards different categories of people. The first latent class indicates a reverse pattern compared to the second latent class. Subjects belonging to the second latent class seem to be closed towards employees and superiors and relatively open towards their colleagues and, to

a lesser extend, customers. The third latent class shows relatively high self-disclosure towards superiors.

Both the variance of the expected category scores as well as the mean intra-person variance of the scores over the self-disclosure variables for each latent class were computed. The variance of the expected category scores over self-disclosure responses for each latent class describes the variance of the response patterns which can be seen in Figure 2.4. This variance is the largest for latent class two (.0022), which is twice as large as for latent class one, which has a variance of .0011. The variance of the expected category scores for latent class three is .0017. This indicates that latent class two is characterized by a response pattern with a relatively large variability over the expected category scores of all latent classes.

The mean intra-person variances (see Equation 2.13) for the three latent classes are .0487, .0246, .0230 respectively. For latent class one, the mean intra-person variance is twice as large as the variances of the other two latent classes. This means that within latent class one, on average subjects have more variance in their response patterns, compared to the other latent classes, and therefore are the most selective in their choice to whom they will show self-disclosure. The mean intra-person variance for subjects belonging to latent class two was relatively low. These subjects are not very selective in their self-disclosure and attain relatively low scores on self-disclosure, even though the variance of the expected category scores for this class was relatively high. The variable pattern of expected category scores shown by this latent class indicates that this latent class is a relatively homogeneous subgroup, consisting of subjects with relatively similar response patterns. It can be concluded that they are closed towards the different categories of people in the work environment.

Latent class one, on the other hand, is a rather heterogeneous subgroup. The subjects belonging to this latent class are relatively selective in their self-disclosure, which means that their degree of self-disclosure depends on the recipient of the self-disclosure. As the variance of the expected category scores for this latent class is relatively low, the response patterns of the subjects belonging to this latent class should be relatively different to each other, to create a flattened pattern of expected category scores of the whole group, with a low variance. The subjects themselves have more variable patterns than is shown by the pattern of their expected category scores showed in Figure 2.4. It

is contrary to the expectations that latent class one consists of subjects with the largest differences in scores of self-disclosure towards the different categories of people. It means that they are relatively selective in their self-disclosure, compared to the other two latent classes, while it was expected that subjects with lower scores on self-disclosure would be the most selective in their self-disclosure.

Latent class three shows a low mean intra-person variance and a variance of the expected category scores which has a value in between those of the other two latent classes. Their expected category scores on the self-disclosure variables also lie in between those of the other two latent classes. Subjects belonging to this latent class are inclined to show a moderate degree of self-disclosure to others, and are not very selective in their choice to whom they will show self-disclosure.

The latent classes differ in their general tendency to self-disclose. Furthermore, the latent classes can be characterized by qualitative different response pattern, which can be interpreted in terms of differences in selectivity in self-disclosure. This supports the first hypothesis. Next, the associations of the latent variables with the collateral variable will be examined.

2.4.2 Associations with Extraversion

The hypothesis concerning the associations between the latent variables and extraversion is tested. The difference between the L^2-statistics of both models yields a test for the associations between extraversion and each of the latent variables, with degrees of freedom equal to the difference of degrees of freedom between both models. The AIC-statistics are given as well. The results of the models analyzed are shown in Table 2.4. It is

Table 2.4: *Results of the Analyses of the Best Fitting Model and the Same Model Without the Association Between Extraversion and Each of the Latent Variables*

Model	L^2	AIC	df
3 - Model with $Y=$ 3 classes	4546.5930	-60715.4070	32631
5 - Model 3 without association between extraversion and θ_Q	4577.9000	-60698.1000	32638
6 - Model 3 without association between extraversion and Y	4612.7725	-60665.2275	32639

seen that the relationship between the latent trait variable and extraversion is significant, $L^2(7) = 31.307$, p<.01. This supports the hypothesis that there is an association between extraversion and the latent trait. The difference between Model 6 and Model 3 is also significant, $L^2(8) = 66.1795$, p<.01. This indicates that there is also an association between extraversion and the latent class variable. The AIC-statistics yield the same results.

Table 2.5 exhibits the category scores, μ_e, of the association between the latent trait variable and extraversion. The association parameter, ϕ^{EQ}, has a value of 29.5226. Extraversion was expected to be a non-decreasing function of the latent trait variable. However, the category scores of Table 2.5 cannot be described as non-decreasing as one scores higher on extraversion.

Table 2.5: *Category Scores, μ_e, of the Association Between the Latent Trait Variable and Extraversion*

Response Category	1	2	3	4	5	6	7	8
Category Score	-0.3257	-0.5841	-0.1416	0.1935	-0.1469	0.4492	0.0354	0.5202

The expected category scores on extraversion in the three latent classes are .0624, -.1257 and .0272 respectively. The latent class showing the highest degree of observed self-disclosure (Latent Class 1) also has the highest expected category score on extraversion, and the latent class showing the lowest observed self-disclosure (Latent Class 2) also shows the lowest expected score on extraversion. This supports the hypothesis that subjects who will show more self-disclosure are more extravert compared to subjects showing less self-disclosure. This association, however, cannot be described as non-decreasing on an individual level (2a).

Finally, Spearman's correlation coefficient is .021 for the relationship between the intra-person variance over self-disclosure variables (selectivity) of the subjects in the sample and their score on extraversion. This association is not significant, p>.05. Thus, the hypothesis that more extravert subjects are less selective in their self-disclosure is not supported (2b).

2.5 Conclusion and Discussion

The results showed that the model with a latent trait variable, and a latent class variable with three categories has the best fit to the data. The four self-disclosure variables appear to be increasing functions of the latent trait, as they should be. The higher one scores on the latent trait, the higher one also scores on the self-disclosure variables.

Next, the relation between the latent class variable and the self-disclosure variables was examined. Subjects in the latent classes differ in their general tendency to self-disclose, as well as in the patterns of the scores on the self-disclosure variables. The differences in patterns could be interpreted in terms of differences in selectivity in self-disclosure, where the first latent class appeared to consist of subjects who are relatively the most selective in their self-disclosure.

When extraversion was examined, the results indicated that both latent variables have an association with extraversion. The relation between the latent trait and extraversion could not be described as a non-decreasing function. The hypothesis that subjects who are the most selective in their self-disclosure would have the lowest scores on extraversion was not supported.

Summarized, the results indicate that qualitative as well as quantitative aspects of self-disclosure can be identified. On top of self-disclosure as a general personality construct, selectivity in self-disclosure appears to play a part in the process of self-disclosure. Subjects in different latent classes differ in their general tendency to self-disclose and have different response patterns of measured self-disclosure. The importance of the notion that it may matter whom someone is facing in deciding whether to show self-disclosure or not, was demonstrated. The aspect of situational specificity may become an extension of the literature on self-disclosure.

One of the first studies on self-disclosure, by Jourard and Lasakow (1958), found differences in self-disclosure towards different target-persons, like towards the mother, father or friends. A study by Slobin, Miller and Porter (1968) on self-disclosure at four organizational levels, indicated that subjects showed the greatest self-disclosure towards their colleagues. Furthermore, Slobin et al. (1968) found more willingness to self-disclose towards superiors than to disclose towards employees. The second latent class, consisting of relatively closed subjects, showed a relatively high degree of self-disclosure towards

colleagues, compared to the other categories of people, which is consistent with the first result of Slobin et al.. Regarding self-disclosure towards superiors, the stated pattern only emerged in latent class three, where self-disclosure towards superiors is even greater than disclosure towards colleagues. A possible explanation for the relatively high self-disclosure willingness towards superiors is that disclosure to a superior may be seen as an ingratiating strategy. This implies that subjects on a lower level in the organizational hierarchy disclose more to people with a higher status, with the hope of reciprocal self-disclosure. This in turn, would equalize their status (Slobin et al., 1968). In the first latent class, a relatively high degree of self-disclosure towards both superiors and employees was observed. This pattern clearly deviates from the findings of Slobin et al. by revealing a relatively high degree of self-disclosure towards employees.

Selectivity in self-disclosure may reflect different motives. Miller and Read (1987), as well as Omarzu (2000), proposed that the decision to self-disclose in a given situation may depend on the goals that an individual wants to attain. For the people belonging to latent class two, who appear to be closed and relatively consistent in their self-disclosure, self-disclosure may not be the preferred strategy to attain their goals. In latent class one on the contrary, self-disclosure is probably seen as a way to attain goals. These differences may also be related to differences in interpersonal trust and the perceived risk of self-disclosure (Steel, 1991). A high perceived risk of self-disclosure may have led the subjects of latent class two to decide not to show self-disclosure. Apart from examining whether people differ in their self-disclosure towards different categories of people, it is also shown that people differ in the degree to which they are selective in their choice to whom they will show self-disclosure. What the reasons are that people differ in the degree of selectivity in self-disclosure, and whether status, trust or the subjective view of the risk of self-disclosure are related to that, is something which may be studied in further research.

A limitation of this study is that self-disclosure was measured towards others who all have to do with the work environment. Although its relationship to the stable personality trait extraversion suggests that the general tendency to self-disclose can be generalized to other situations, the patterns of self-disclosure are specific to the work environment. People may show relatively little self-disclosure in their work environment, but might respond differently in other environments, like home. In a home situation, where people

can feel safe and free, it can be expected that people will have less difficulty to show self-disclosure. They are among the people closest to them. Measures of self-disclosure towards other categories of people could provide more insight in this issue.

Mixture measurement modeling provided a useful method to study quantitative differences, differences in the general degree of self-disclosure, as well as qualitative differences in self-disclosure towards the different categories of people. In this way, application of a mixture measurement model contributes to research on construct validation. Furthermore, it may lead to a better prediction of disclosure behavior in general as well as in specific classes of situations. It is possible to examine whether, apart from a general personality construct, different patterns of responses can be distinguished.

The proposed latent-trait latent-class model can be used to allocate subjects to the different latent classes and to determine their score on the discrete latent trait. From sample data, the probability of belonging to each of the latent classes and attaining a certain score on the latent trait, can be estimated from the subjects' manifest responses. The class someone belongs to is usually chosen as the class with the highest conditional probability given his or her manifest response pattern. The probability of each combination of the latent variables, given the responses x_1 through x_4, is

$$\pi_{\theta_q m|x_1x_2x_3x_4} = \frac{\pi_{\theta_q}\pi_m\pi_{x_1|\theta_q m}\pi_{x_2|\theta_q m}\pi_{x_3|\theta_q m}\pi_{x_4|\theta_q m}}{\sum_{\theta_q}\sum_m \pi_{\theta_q}\pi_m\pi_{x_1|\theta_q m}\pi_{x_2|\theta_q m}\pi_{x_3|\theta_q m}\pi_{x_4|\theta_q m}}.$$

Based on the conditional probabilities of the manifest variables given the values of the latent variables, and the probabilities of the latent variables itself, the probability can be computed of each combination of the latent variables, given the manifest responses. In this way it is possible to determine both the subject's degree and type of selectivity in his self-disclosure.

Self-disclosure is a difficult concept, because it involves both qualitative and quantitative aspects. This research showed that it is possible to identify both aspects. It may help in providing us with a better understanding of self-disclosure, and when and why people differ in their degree of self-disclosure towards different people. This research also illustrates the possible use of latent-trait latent-class models for the analysis of differences in item responses between subjects.

Chapter 3

Improvement in Detection of Differential Item Functioning Using a Mixture Item Response Theory Model

Usually, methods for detection of differential item functioning (DIF) compare the functioning of items across manifest groups. However, the manifest groups with respect to which the items function differentially may not necessarily coincide with the true source of bias. It is expected that DIF detection under a model that includes a latent DIF variable is more sensitive to this true source of bias. In a simulation study, it is shown that a mixture item response theory model, which includes a latent grouping variable, performs better in identifying DIF items than DIF detection methods using manifest variables only. The difference between manifest and latent DIF detection increases as the correlation between the manifest variable and the true source of the DIF becomes smaller.

An item is said to exhibit differential item functioning (DIF) when it has different statistical properties for different groups, after matching the groups with respect to ability (Angoff, 1993). A number of methods have been proposed to examine DIF (Camilli & Shepard, 1994; Holland & Wainer, 1993). Modern IRT based DIF methods define DIF as the presence of differences in the probability of a correct response for two manifest groups, conditional on the latent trait. The differences in the conditional probabilities are reflected

This chapter has been submitted for publication.

by differences in item parameters across manifest groups, which can be graphically represented by differences in the item characteristic curves (ICCs) of the groups. Differential item functioning can be quantified by estimating the area between the ICCs for the two groups (Raju, 1988) or improvement in model fit by comparing a model with and without a grouping parameter (Thissen, Steinberg, & Wainer, 1993). Furthermore, DIF can be assessed by comparing the item parameters across groups (Lord, 1980).

These DIF detection methods compare the functioning of an item across manifest groups. It is assumed that the manifest groups, for example males and females or ethnic groups, represent homogeneous subgroups for which the items function the same way, that is, the items do not exhibit DIF within the manifest groups. Furthermore, it is assumed that these manifest variables are related to the source of the DIF. DIF analyses with a mixture Rasch model show that the manifest groups used for DIF detection are not necessarily identical to the groups of subjects for whom an item may be biased (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005). Mixture item response theory (IRT) models (Kelderman & Macready, 1990; Mislevy & Verhels, 1990; Rost, 1990; 1991) make no assumptions as to the type or cause of qualitative differences in the responses of subjects. Homogeneous subgroups, latent classes, are identified where within each latent class the same measurement model holds but with different parameter estimates across latent classes. Let X_{ij} denote a response by subject i to item $j \in \{1, \ldots, J\}$, with realizations $x_{ij} = k \in \{1, \ldots, K_j\}$. Furthermore, let Y_i denote latent class membership of subject i with realizations $y_i = m \in \{1, \ldots, M\}$, and let θ_{mi} denote the latent trait value of a subject in latent class m. For a mixture of the two-parameter logistic (2PL, Smit, Kelderman, & Van der Flier, 2000) model, the probability that a subject i from latent class m with latent trait value θ_{mi} gives a correct response to item j is formulated as,

$$P(X_{ij} = 1|\theta_{mi}, y_i = m) = \frac{\exp(\alpha_{jm}(\theta_{mi} - \delta_{jm}))}{1 + \exp(\alpha_{jm}(\theta_{mi} - \delta_{jm}))} \tag{3.1}$$

where the item discrimination α_{jm} and difficulty parameters δ_{jm} may vary across latent classes m. Similarly, for manifest DIF detection methods the item parameters may vary across manifest groups. However, mixture IRT models should be more suitable to identify the true source of the DIF, because they are not restricted to detect DIF relative to specific manifest variables. Nevertheless, one or more exogenous, manifest variables may be incorporated in mixture IRT models as collateral variables (Smit, Kelderman, & Van

der Flier, 1999; Smit et al., 2000). Incorporating manifest variables reduces the standard errors of the item parameter estimates and may aid in the interpretation of the latent classes (Smit et al., 2000).

Kelderman and Macready (1990) were the first to use mixture IRT models to assess differential item functioning. They compared DIF detection across manifest and latent groups. Since then, several studies have demonstrated that the latent classes may show clear differences compared to the manifest groups that were assumed to be related to the source of the DIF (e.g. Cohen et al., 2005; Zickar, Gibby, & Robie, 2004). Cohen and Bolt (2005) showed that manifest characteristics associated with DIF, like gender, often had a weak relationship with the latent classes. They used a mixture Rasch model to identify biased items in a mathematics placement test. Three latent classes were identified, where the nature of the mathematics problems made the items differentially easier or harder for one class compared to the others. It was shown that mixture IRT models allow for detection of DIF, while also allowing for direct interpretation of the possible causes of DIF.

De Ayala, Kim, Stapleton and Dayton (2002) compared several manifest DIF detection methods with respect to their performance of identifying DIF items in data that were based on a mixture IRT model. It was shown that as the association between the latent classes and the manifest variable was reduced, the performance of the manifest DIF detection methods decreased rapidly. However, it was not examined how a mixture IRT model would perform in identifying the biased items compared to the manifest DIF detection methods. Von Davier and Yamamoto (2004) analyzed a mixture IRT model where class membership information was partially observed. Again, this approach requires strong assumptions as to which variables correctly represent the latent classes, and thus may be associated with DIF.

Mixture IRT models for detection of DIF differ from the manifest DIF detection methods with respect to the nature of the grouping variable. Differences in item parameters may be studied across latent or manifest groups. For both methods, the equality of the item parameters across groups are investigated. Lord (1980) described a chi-square method for comparing vectors of item parameters for two groups of subjects, say $m = 1$ and $m = 2$. Let $\mathbf{z}_j'$ be a vector $\{\hat{\delta}_{j1} - \hat{\delta}_{j2}, \hat{\alpha}_{j1} - \hat{\alpha}_{j2}\}$ and $\hat{\sum}_j$ be the asymptotic variance/covariance matrix for $\mathbf{z}_j$. The χ_j^2 statistic in Equation 3.2 tests whether both

the discrimination and difficulty parameters are equal across groups, and is chi-square distributed with degrees of freedom equal to the number of parameters being tested.

$$\chi_j^2 = \mathbf{z}_j' \hat{\sum}_j^{-1} \mathbf{z}_j \tag{3.2}$$

It can be examined for which items the parameters vary significantly across the manifest or latent groups, and thus are identified as displaying DIF.

When the discrimination parameters are restricted to equal 1, one can test for uniform DIF. The differences between the difficulty parameters across two groups can be examined after Equation 3.2 has been reduced to,

$$\chi_j^2 = \frac{(\hat{\delta}_{j1} - \hat{\delta}_{j2})^2}{V_j} \tag{3.3}$$

where V_j denotes the variance of the differences in difficulty parameters of the two groups (Lord, 1980; Thissen et al., 1993). For two manifest groups, V_j is equal to the sum of the variances of the two difficulty parameters. This statistic is asymptotically chi-square distributed with one degree of freedom. When the χ_j^2 statistic exceeds the critical value for a given level of significance, an item is said to exhibit DIF. Note, that the item parameters need to be on a common scale before comparisons across groups can be made. The difficulty parameters may be scaled by fixing the mean $\bar{\delta}'_m$ of the scaled difficulty parameters $\hat{\delta}'_{jm}$ in each group m equal to zero;

$$\hat{\delta}'_{jm} = \hat{\delta}_{jm} - \bar{\delta}_m. \tag{3.4}$$

When there are larger numbers of items, it is recommended to perform the DIF detection iteratively. The item with the largest χ_j^2 value is designated as the first item that exhibits DIF. This item is removed, and the scaling of the difficulty parameters is performed on the remaining set of items. After the scaling, the test statistics are examined to identify the next item exhibiting DIF. The scaling is performed again on the remaining items, and so on until no more item can be identified as displaying DIF, or no additional information can be obtained.

In this study, it will be examined whether mixture IRT models, with a latent grouping variable, perform better in identifying DIF items compared to DIF detection methods using manifest variables only. A simulation study is conducted to examine the ability to correctly identify DIF items for the DIF detection methods with manifest and

with latent groups. It has been shown that the association between the latent classes and the manifest variable influences the identification of DIF items (De Ayala et al., 2002). Therefore, the performance of the two methods will be examined under different conditions, varying in the strength of the association between the latent classes and the manifest, exogenous variable. It is expected that the manifest DIF detection method will perform worse as the association is reduced. The DIF detection method for latent groups is expected to perform relatively well in identifying DIF items, irrespective of the strength of the association. Still, it is expected that the identification of DIF items is better when there is a strong compared to a weak association between the latent class and manifest variable. Also, a mixture IRT model that does not include the manifest variable will be examined. This allows us to examine the influence of the manifest variable on the identification of the latent classes, when there is no association. When differences are observed, conditions for including manifest variables in mixture IRT models can be made more specific. Sample size is also included as a design variable, where it is expected that a larger sample size will lead to a better identification of DIF items.

3.1 Simulation study

A manifest and a latent DIF detection method will be compared with respect to their ability to identify DIF items. It will be examined how many DIF items are not identified as displaying DIF, that is, what the number and proportion of false negatives is (Type II error). Also, the number and proportion of false positive identifications will be examined. This is the Type I error, which concerns the number of non-DIF items that are incorrectly identified as displaying DIF.

3.1.1 Design

Data sets of 5000 and of 25000 subjects are generated according to a mixture version of the 2PL model as shown in Equation 3.1, where the discrimination parameters are restricted to equal 1. There are two latent classes of equal size. For both latent classes, the latent trait is sampled from a standard normal distribution. A fixed test length of 27 items was used. For the first latent class, the difficulty parameters correspond to a truncated standard normal distribution from -1 through 1. Nine items were simulated to

display DIF in the difficulty parameters. This resulted in increased parameter values for the second latent class. The degree of DIF varies across the DIF items, ranging from 0.3 through 1.1, with increases of 0.1. This allows us to examine under which DIF conditions each method is effective in identifying DIF items. It is expected that the items with the largest DIF will have the largest probability to be correctly identified. Furthermore, it is expected that DIF detection methods including the latent grouping variable will be more effective in identifying small DIF in the items, compared to the DIF detection based only on the manifest variable.

The set of 27 items is divided into three sets of nine subsequent items, where within each set an item will display a small, medium and large degree of DIF. The second, fifth and eight item of each set will display DIF, to equally distribute the DIF across the test. The last set of nine items has the highest difficulty parameters. To prevent extreme values, the smaller DIF (values of 0.3, 0.6 and 0.9) will be allocated to these items. This means that item 26, the DIF item with the highest difficulty parameter, will display the smallest amount of DIF, that is 0.3. Item 20 will display the largest DIF, which leaves Item 23 to display a medium degree of DIF. This results in a maximum value of the difficulty parameter of 1.3 for item 20. The same distribution of DIF across the nine items will be followed for the first and second set of nine items. The resulting difficulty parameters that will be used in the simulation study, as well as the degree of DIF, are given in Table 3.1.

Table 3.1: *Difficulty Parameters δ_{jm} of the Simulation Study*

Item	1	2	3	4	5	6	7	8	9
Latent class 1	-0.99	-0.89	-0.80	-0.71	-0.63	-0.55	-0.48	-0.40	-0.33
Latent class 2	-0.99	0.21	-0.80	-0.71	0.17	-0.55	-0.48	0.10	-0.33
DIF		1.1			0.8			0.5	
Item	10	11	12	13	14	15	16	17	18
Latent class 1	-0.27	-0.20	-0.13	-0.07	0	0.07	0.13	0.20	0.27
Latent class 2	-0.27	0.80	-0.13	-0.07	0.70	0.07	0.13	0.60	0.27
DIF		1.0			0.7			0.4	
Item	19	20	21	22	23	24	25	26	27
Latent class 1	0.33	0.40	0.48	0.55	0.63	0.71	0.80	0.89	0.99
Latent class 2	0.33	1.30	0.48	0.55	1.23	0.71	0.80	1.19	0.99
DIF		0.9			0.6			0.3	

The strength of the association between the simulated latent class variable, the

source of the bias, and the manifest variable influences the identification of DIF items (De Ayala et al., 2002). Therefore, six conditions are examined, differing in the strength of the association between the simulated latent classes and the manifest variable. Furthermore, it will be examined whether a mixture IRT model including the latent grouping variable with a zero correlation between the simulated latent class and manifest variable produces comparable results as a mixture IRT model excluding the exogenous variable. The manifest variable is a dichotomous variable, with varying overlap with simulated latent class membership. Sample size is included as a design variable as well. An overview of the design of the simulation study is given in Table 3.2.

Table 3.2: *Design of the Simulation Study*

	Correlation					
Grouping Variable	0	0.2	0.4	0.6	0.8	1
Sample size of 5000						
Manifest	a1	b1	c1	d1	e1	g1
Latent, with manifest indicator	a2	b2	c2	d2	e2	g2
Latent	a3	-	-	-	-	-
Sample size of 25000						
Manifest	A1	B1	C1	D1	E1	G1
Latent, with manifest indicator	A2	B2	C2	D2	E2	G2
Latent	A3	-	-	-	-	-

The data generation for each condition is replicated ten times. For each replicated data set, both methods will be used to assess DIF. For the manifest DIF detection method, maximum likelihood estimates of the item parameters and the corresponding variance/ covariance matrix are obtained by fitting the mixture 2PL model to the data (Vermunt, 1997). The discrimination parameters are restricted to equal 1, and the difficulty parameters may vary across the observed groups. Next, the mixture 2PL model is fitted, where the grouping variable is unobserved, but may be related to the manifest variable. This mixture IRT model with two latent classes will be fitted five times with random starting values to avoid the identification of local maxima. For Condition A, with a zero correlation between the manifest and latent variable, also the mixture 2PL model with two latent classes is fitted excluding the manifest, exogenous variable.

Next, the maximum likelihood estimates of the difficulty parameters of the manifest

or latent groups need to be placed on a common scale. The mean of the scaled difficulty parameters $\overline{\delta}'_m$ for each group m is fixed to 0. This is obtained by subtracting the mean of group m from each difficulty parameter value δ_{jm}, as shown in Equation 3.4. The same transformation is performed on the variance/covariance matrix of the item parameters. An item will be marked as exhibiting DIF when the test statistic given in Equation 3.3 exceeds the critical value. The critical value is based on the chi-squared distribution with one degree of freedom, and will be inspected for $\alpha = .05$ and $\alpha = .01$. Because there is quite a large number of items J, also a Bonferonni-correction of the significance level will be inspected, where $\alpha = .05/J$. The identification of items exhibiting DIF will be performed iteratively, as described above.

3.1.2 Results

For each method, condition and replication, the number of times an item is identified as exhibiting DIF is recorded. The number of times each item was identified as displaying DIF summed over the number of replications are given in Table 3.3, for the conditions with a sample size of 5000 subjects and $\alpha = .05$. Also, the total number and proportion of non-DIF items that are incorrectly identified as displaying DIF (false positives, FP) and of DIF items that are not identified as exhibiting DIF (false negatives, FN) are given. First, focus on the results of DIF detection using the manifest variable only. When the manifest variable is unrelated to the source of bias, the manifest DIF detection method performs poorly. The number of false positive identifications is larger than the number of items correctly identified as displaying DIF. The number of correct identifications increases as the correlation of the manifest variable with the simulated latent classes increases. The number of false negatives (FN) is reduced to zero with increasing correlations. A strong correlation of 0.8 or 1 between the manifest and simulated latent variable leads to perfect identification of the DIF items. Still, in each condition a number of non-DIF items are falsely identified as displaying DIF.

The DIF detection with the latent grouping variable performs better in identifying DIF items compared to the manifest grouping variable, in particular when the correlation between the manifest and simulated latent variable is low. The number of FN is much lower for the conditions with correlations of 0 through 0.4. The two procedures perform equally well, with regard to the number of FN, for the conditions with correlations of 0.6

Table 3.3: *Number of Times Each Item Was Identified as Exhibiting DIF, for Conditions with Sample Size 5000 and* $\alpha = .05$*, as well as the Total Number (♯) and Proportions (%) of False Positives and False Negatives*

	Correlation simulated latent class and manifest variable												
	0		0.2		0.4		0.6		0.8		1		
Item	Lat†	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat
1	0	0	1	1	1	1	0	1	0	1	1	0	0
2 *	9	0	9	9	10	10	10	10	10	10	10	10	10
3	0	1	0	0	0	0	1	1	1	2	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0
5 *	10	2	10	4	10	10	10	10	10	10	10	10	10
6	1	0	1	1	0	1	1	1	1	1	1	0	0
7	1	1	2	1	0	1	1	0	0	2	2	0	0
8 *	5	0	5	3	6	10	10	10	10	10	10	10	10
9	1	1	0	0	0	0	1	1	1	2	2	1	1
10	1	0	1	2	0	1	0	1	1	1	1	0	0
11 *	10	1	9	9	10	10	10	10	10	10	10	10	10
12	1	1	1	1	0	0	0	0	0	0	0	1	1
13	1	1	0	0	1	1	2	0	0	0	0	1	1
14 *	10	0	7	8	10	10	10	10	10	10	10	10	10
15	1	1	0	1	1	0	0	0	1	0	0	0	0
16	1	1	1	2	0	0	0	1	1	1	0	0	0
17 *	6	0	4	0	7	8	9	10	10	10	10	10	10
18	0	1	0	1	0	2	0	0	1	0	0	1	1
19	0	1	0	0	0	1	3	1	1	0	1	0	0
20 *	10	1	7	3	10	10	10	10	10	10	10	10	10
21	1	2	0	0	1	2	0	0	0	0	0	0	0
22	0	0	1	0	1	0	0	2	2	1	1	2	2
23 *	6	2	4	3	9	10	10	10	10	10	10	10	10
24	0	1	0	0	0	1	1	2	4	0	2	0	0
25	2	0	1	0	1	1	0	1	0	0	1	0	0
26 *	3	0	1	2	7	2	5	9	9	10	10	10	10
27	0	0	0	1	0	0	0	1	0	0	0	0	0
Total	80	18	66	52	85	92	94	102	103	101	102	96	96
♯ FP	11	12	10	11	6	12	10	13	14	11	12	6	6
% FP	6	7	6	6	3	7	6	7	8	6	7	3	3
♯ FN	21	84	34	49	11	10	6	1	1	0	0	0	0
% FN	23	93	38	54	12	11	7	1	1	0	0	0	0

Note. FP = False Positives, FN = False Negatives; DIF detection based on:
Lat† = latent grouping variable, for mixture IRT model excluding manifest variable,
Man = manifest grouping variable, Lat = latent grouping variable.
* denotes DIF item

or higher. The results with respect to the number of FP are comparable to those of the manifest DIF detection procedure. These results suggest that in particular in situations where the manifest variable may be a bad representation of the variable truly associated with DIF, a latent grouping variable may provide better results in correctly identifying DIF items.

It was examined whether the manifest variable incorporated in the mixture IRT model interferes with the identification of latent classes corresponding to the true source of the DIF. This could weaken the identification of DIF with a latent grouping variable. In that case, the mixture IRT model excluding the manifest variable would produce better results in identifying DIF items. The results of this model are given in the second column of Table 3.3, denoted by Lat†. The number of FN is obviously lower, compared to the DIF detection with the latent grouping variable based on the mixture IRT model including the manifest variable. However, when there is even a very low correlation between the manifest and simulated latent variable, including the manifest variable in the mixture IRT model should be preferred. The number of FN is halved for the condition with correlation 0.2, compared to the results of the mixture IRT model excluding the manifest variable.

It is seen in Table 3.3 that the items with the smallest DIF, Items 8, 17 and 26, are more difficult to identify compared to the other DIF items. The identification improves as the correlation increases. Item 26 is only identified in all replications when the correlation between the manifest and simulated latent variable is 0.8 or higher. This pattern can be seen for the DIF identification for both the manifest and latent groups.

The critical value of the test statistic is based on the χ^2 distribution where the level of significance was set at .05. Because there is quite a large number of items, there is a high probability of identifying some items as displaying DIF, even when there is no DIF. Therefore, two more decision criteria are employed as well, a significance level of .01 and a Bonferonni-correction resulting in $\alpha = .05/J$, where J is the number of items that is examined. The total number and proportions of FP and FN for the DIF detection in each condition with samples size of 5000 subjects, based on these restricted α levels are given in Table 3.4.

The number of items identified as displaying DIF decreases when a more restricted level of significance is imposed. Consequently, the number of FP decreases, for all procedures. The zero correlation condition shows slightly more FP identifications for

Table 3.4: *Number (♯) and Proportions (%) of False Positives and False Negatives, for Conditions with Sample Size 5000 and Two Levels of Significance ($\alpha = .01$ and $\alpha = .05/J$)*

	Correlation simulated latent class and manifest variable												
	0		0.2		0.4		0.6		0.8		1		
	Lat†	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat
						$\alpha = .01$							
♯ FP	7	2	6	6	2	1	1	3	4	1	2	0	0
% FP	4	1	3	3	1	1	1	2	2	1	1	0	0
♯ FN	31	90	44	66	19	18	11	8	8	2	0	0	0
% FN	34	100	49	73	21	20	12	9	9	2	0	0	0
						$\alpha = .05/J$							
♯ FP	6	0	5	3	1	0	0	0	1	0	0	0	0
% FP	3	0	3	2	1	0	0	0	1	0	0	0	0
♯ FN	39	90	57	78	33	27	18	14	11	2	4	0	1
% FN	43	100	63	87	37	30	20	16	12	2	4	0	1

Note. FP = False Positives, FN = False Negatives; DIF detection based on:
Lat† = latent grouping variable, for mixture IRT model excluding manifest variable,
Man = manifest grouping variable, Lat = latent grouping variable.

DIF detection with a latent grouping variable compared to the manifest DIF detection, because the manifest DIF detection method hardly identifies any item as displaying DIF. The DIF detection method with a latent grouping variable performs a lot better than with a manifest grouping variable, the number of FN is much lower. Naturally, the number of FN increases when a more restricted level of significance is imposed. As the correlation between the manifest and simulated latent variable increases, the differences in the number of FN for DIF detection methods with manifest and latent grouping variables are minimized. When lower correlations between the manifest variable and the source of the DIF may be expected, DIF detection with a latent grouping variable provides better results in identifying DIF items compared to a manifest DIF detection method. For the condition with a zero correlation, the mixture IRT model excluding the manifest variable is again more effective in identifying DIF items as displaying DIF compared to the other methods. The number of FN for Lat† is lower compared to the methods including the manifest variable. The number of FP identifications is comparable for the two mixture IRT models in the condition with a zero correlation.

Table 3.5: *Number (♯) and Proportions (%) of False Positives and False Negatives, for the Conditions with Sample Size 25000 and Different Levels of Significance*

	Correlation simulated latent class and manifest variable												
	0		0.2		0.4		0.6		0.8		1		
	Lat†	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat	Man	Lat
$\alpha = .05$													
♯ FP	7	4	8	7	12	15	15	8	8	5	5	8	2
% FP	4	2	4	4	7	8	8	4	4	3	3	4	1
♯ FN	0	84	0	4	0	0	0	0	0	0	0	0	0
% FN	0	93	0	4	0	0	0	0	0	0	0	0	0
$\alpha = .01$													
♯ FP	2	1	2	1	3	5	1	1	0	0	1	0	0
% FP	1	1	1	1	2	3	1	1	0	0	1	0	0
♯ FN	0	88	1	13	0	2	0	0	0	0	0	0	0
% FN	0	98	1	14	0	2	0	0	0	0	0	0	0
$\alpha = .05/J$													
♯ FP	0	0	0	0	1	3	1	0	0	0	0	0	0
% FP	0	0	0	0	1	2	1	0	0	0	0	0	0
♯ FN	0	90	1	21	1	2	0	0	0	0	0	0	0
% FN	0	100	1	23	1	2	0	0	0	0	0	0	0

Note. FP = False Positives, FN = False Negatives; DIF detection based on:
Lat† = latent grouping variable, for mixture IRT model excluding manifest variable,
Man = manifest grouping variable, Lat = latent grouping variable.

These conditions and significance levels are also examined for data sets with a larger sample size, in this case samples of 25000 subjects. The total number and proportions of FP and FN based on the three levels of significance, are given in Table 3.5. A large decrease in the number and proportion of FN is seen compared to the smaller sample size. Only the conditions with correlations of 0 and 0.2 show a large number of FN for the manifest DIF detection method. In these conditions, DIF detection is nearly perfect for both mixture IRT models with and without the manifest variable. However, for less restricted levels of significance there are still a number of FP identifications. Finally, it is seen that the inclusion of a manifest variable unrelated to the true source of DIF does not interfere with the identification of the latent classes.

3.2 Conclusion

This study focused on the identification of items that function differentially across groups. A well-known DIF detection method compares IRT parameters across manifest groups using a χ_j^2 statistic (Lord, 1980), see Equation 3.3. A similar statistic is based on the comparison of IRT parameters of a mixture Rasch model (Kelderman & Macready, 1990; Rost, 1990), where the grouping variable is latent. The latent classes are based on the response patterns and supposed to be related to the source of the bias. A simulation study was performed to compare the effectiveness in identifying DIF items of the manifest and latent DIF detection methods. It can be concluded that the mixture IRT model performs better in identifying DIF items compared to DIF detection using manifest variables only. The differences between the two methods increase as the correlation between the manifest variable and the source of the DIF becomes smaller. This is in agreement with the results of De Ayala et al. (2002). They showed that as the association between the source of the bias and the manifest variable was reduced, the performance of several manifest DIF detection methods decreased rapidly. However, they did not investigate a mixture IRT model to identify DIF items.

It has been shown that even when there is a small correlation between the source of bias and the manifest variable, including this manifest variable as an indicator of the latent class variable improves the identification of DIF. The manifest and latent DIF detection methods perform better when the estimation of the item parameters is based on a larger sample size. However, the manifest DIF detection method is not very efficient in detecting DIF in conditions with a low correlation between the manifest variable and the source of bias.

The present study tests for uniform DIF, that is, only the item difficulty parameters were allowed to vary across latent classes. This could be extended to non-uniform DIF, by allowing item discrimination parameters to vary across latent classes as well. The χ_j^2 statistic shown in Equation 3.2 can be used to identify items exhibiting non-uniform DIF. Furthermore, this statistic can be generalized to analyze more than two sets of item parameters. A multigroup chi-squared statistic has been developed (Kim, Cohen, & Park, 1995) to analyze DIF across multiple groups simultaneously, instead of making pairwise comparisons. The multigroup statistic can also be used to study DIF across multiple

latent classes simultaneously.

The latent DIF detection method need not restrict itself to the identification of DIF associated with a specific manifest variable. It allows us to examine the true source of the DIF. Nonetheless, manifest variables may be incorporated as indicators of the latent class variable to study their association. Future research may be directed at incorporation of more than one manifest variable as indicator of latent class membership, to improve the identification of DIF items.

In this study, it was assumed that there are several biased items. It was shown that the latent DIF detection method has a surplus value when there are a number of DIF items to identify. When there is just one item displaying DIF, it would be virtually impossible to identify the differential functioning of this item with a pure mixture IRT model excluding exogenous variables. In that case, adding a manifest indicator variable would render the latent class variable superfluous, since there is only a single path from the manifest variable to the item response. However, there are usually several biased items. When the manifest variable is not an entirely valid indicator of the true source of DIF, the latent grouping variable would help in identifying this common bias. We have seen in the simulation study that even if the validity of the manifest variable is high, including a latent grouping variable does no harm. Future research could study the identification of DIF items when there are only a few biased items. This could be examined for different degrees of bias. In this way, conditions for applying latent as opposed to manifest DIF detection methods could be made more specific.

In summary, it can be concluded that mixture IRT models are a good alternative for DIF identification compared to manifest DIF detection methods. In particular, in situations where the origin or source of the differential functioning of the items is not evident, which is quite often the case. It provides room to detect the true source of the DIF, without being restricted to specific manifest variables. Even when the manifest grouping variable is a valid indicator of this source, it may be measured with error. For example, in detecting DIF related to cultural differences. The mixture IRT models describe both the within and between group heterogeneity, and in this way may identify homogeneous subgroups for which some items may show differential functioning associated with the true source of the DIF.

Chapter 4

Fitting a Mixture Item Response Theory Model to Personality Questionnaire Data: Characterizing Latent Classes and Investigating Possibilities for Improving Prediction

Mixture IRT models aid the interpretation of the response behavior of subjects on personality tests and may provide possibilities for improving prediction. The heterogeneity in the population is modeled by identifying homogeneous subgroups that conform to different measurement models. Qualitative group differences are only taken into account to the extent that they may be determined from the response patterns. No assumptions as to the type of qualitative differences have to be made beforehand. In this study, mixture IRT models were applied to the extraversion and neuroticism scale of the Amsterdam Biographical Questionnaire (based on Eysenck's Maudsley Personality Inventory). For both scales, a mixture version of the nominal response model with three latent classes was identified as the best fitting model. The latent classes differed with respect to social desirability and ethnic background. Furthermore, the response tendencies within the latent classes demonstrated a differential use of the "?" category. An important issue is whether applying mixture IRT models results in a better prediction of relevant external criteria,

This chapter will be published as: Maij-de Meij, A.M., Kelderman, H., & Van der Flier H. Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement* (in press).

compared to the prediction with a one class model. For the neuroticism scale the prediction improved, but not for the extraversion scale. The results demonstrate the possible advantage of applying mixture IRT models to personality questionnaires.

Psychological measurement and prediction procedures usually assume that individual differences can be described by a single standard measurement and prediction model. However, empirical research has demonstrated that these models do not always hold for all subjects (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990; 1991). Bias research has shown that there are qualitative differences between groups of subjects, differing in gender, ethnic background or other variables. Questionnaires may also be contaminated by response biases that differ across subjects, in particular acquiescence (tendency to agree or disagree) or preference for the middle category, and social desirability.

4.1 Mixture IRT Models

Mixture item response theory (mixture IRT) models have been proposed, to handle the problem of qualitative group differences (Rost, 1990; 1991). The heterogeneity in the population is modeled by identifying homogeneous subgroups, latent classes, that conform to different measurement models. Often, this means that the same measurement model holds within each latent class, but with different parameter estimates across latent classes. When the response behavior can be described by a one class model, the items function the same way for all subjects. However, if subjects use the response scale in different ways, due to whatever cause, more than one latent class will be identified.

Studies of differential item functioning (DIF) use a priori known grouping information. The sample of subjects is split by means of manifest criteria (e.g., gender, ethnic background, or age), and it is analyzed whether the parameters of a specific IRT model are invariant across these groups (Schmitt, Holland, & Dorans, 1993; Thissen, Steinberg, & Wainer, 1993). Mixture IRT models have a priori unknown grouping. Qualitative differences are taken into account to the extent that they may be determined from the response patterns, and do not rely exclusively on information about group membership.

Kelderman and Macready (1990) presented a log-linear Rasch model with a latent

grouping variable to assess differential item functioning across latent classes. Mixture IRT models can provide more information about the cause of DIF, which can, for example, be related to the nature of a mathematical problem (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005). An important issue is that no assumptions have to be made as to which manifest characteristics are associated with the latent classes. Cohen and Bolt (2005) showed that manifest characteristics associated with DIF are not necessarily identical to the characteristics of the latent classes for which an item may be biased.

Often, mixture IRT models focus on differences in the use of the response scale across latent classes, that applies to all items. Rost, Carstensen and Von Davier (1997) analyzed personality scales and identified two latent classes with different response sets. One latent class showed a tendency towards extreme ratings, whereas the other latent class was characterized by moderate ratings. Similar results were found in responses to a job satisfaction survey (Eid & Rauber, 2000). Other applications of mixture IRT models are performed, for example, to identify solution strategies (Mislevy & Verhelst, 1990), examine the effects of test speededness (Bolt, Cohen, & Wollack, 2002), or identify latent classes reflecting situation specific self-disclosure (Maij-de Meij, Kelderman, & Van der Flier, 2005).

4.1.1 Prediction

An important issue, however, that has not gained much attention but is of great importance to psychological applications, is whether applying these mixture IRT models provides possibilities to improve the prediction of external criteria, compared to the same model consisting of one latent class. A mixture IRT model describes the response patterns of subjects in more detail, compared to the one class model, where all subjects are assumed to respond according to the same measurement model. One can imagine that a given measurement model describes the response pattern of a part of the sample reasonably well, providing a good estimate of their latent trait value. However, another part of the sample may use the response scale in quite a different manner, which means that the measurement model does not fit their response behavior and will provide a poor estimate of their latent trait value. Therefore, it can be expected that latent trait scores estimated using a mixture IRT model may provide a better prediction of an external criterion, compared to the latent trait scores estimated with a one class model.

4.1.2 Response Tendencies

The differential use of the response scale can be manifested in different ways. Directed at personality self-report questionnaires, a number of studies have examined the functioning of the middle response category. It has been shown that, for example, the "?" category is interpreted in various ways. The meaning of the "?" category may vary from ambivalent or indifferent, to regarding it as an intermediate category. Some subjects choose it because they refuse to reveal their personal feelings, feel not competent enough to take a position, or do not understand the statement (DuBois & Burns, 1975). Smit, Kelderman, and Van der Flier (2003) applied a mixture version of the nominal response model and the generalized partial credit model to examine two personality scales with response categories "yes", "?" and "no". They indicated that the parameters of the "?" response category were not invariant over subjects. A group of subjects was identified that tended to avoid the "?" category. Similar results were found by Hernández, Drasgow, and González-Romá (2004).

The most extensively studied type of response bias is the tendency to give a socially desirable answer. Paulhus (1984; 1986) distinguishes two aspects of social desirability; impression management and self-deception. Impression management refers to the conscious and deliberate response distortion in situations where it is desirable to present oneself in a positive light. In contrast, self-deception represents a more stable view of oneself in positive terms. Personnel selection situations are settings where subjects want to make a positive impression. The subjects may distort their responses in order to look good and create a favorable outcome (Barrick & Mount, 1996; Bass, 1957; Dunnette, MacCartney, Carlson, & Kirchner, 1962). Dunnett, Koun and Barber (1981), for example, studied the Eysenck Personality Inventory and found it susceptible to faking. Situations where subjects were motivated to give a good or bad impression resulted in biased responses. Furthermore, their study showed that extraversion was perceived as socially desirable and neuroticism as undesirable.

Zickar, Gibby, and Robie (2004) applied the mixture Rasch model to uncover faking subgroups among a sample of applicants and incumbents. Often, applicants and incumbents are used to study faking, where applicants are expected to fake and incumbents to respond honestly. Zickar et al. (2004) identified latent classes differing in the degree

of faking, from honest responding to extreme faking. Furthermore, they found a large overlap in the distribution of applicants and incumbents among the latent classes. There was a sizeable number of applicants appearing to respond honestly and incumbents who appeared to fake.

Another variable that may affect the response behavior is ethnic background. Te Nijenhuis and Van der Flier (1999) reviewed research on test bias against immigrant children and job applicants in The Netherlands. Overall, it was concluded that tests can be used for comparisons between immigrants and majority group members. However, immigrants did show a less favorable score profile for job applicants on several personality inventories. Van de Vijver and Phalet (2004) pointed out the importance of the role of acculturation in assessment in multicultural groups. Usually, there is considerable heterogeneity within cultural groups. It is expected that the heterogeneity in cultural groups will increase as a result of differences in acculturation. Mixture IRT models can describe both within and between-group heterogeneity. The previously described study of Smit, Kelderman and Van der Flier (2003) also contained ethnic background in the mixture IRT models as an exogenous variable. The outcomes suggested that cultural differences in conventions about the disclosure of personal information are associated with differences in item scores.

Response tendencies and other variables influencing the response behavior, may affect the precision of the prediction. Ghiselli (1956; 1963) studied precision of prediction by looking at differential predictability. He argued that the precision of the prediction based on a test score may vary over subjects. A procedure was proposed to differentiate between subjects for whom a test is a good predictor and those for whom it is a poor predictor. A different approach was taken by Bem and Allen (1974). They demonstrated that the predictive power of a test depends on the consistency with which the subjects rate their own behavior. Instead of predictability, they talk about cross-situational consistency. Furthermore, they refer to scalability, which demands that subjects must all scale behaviors in the same way. Goodman (1975) proposed a model to identify "intrinsically scalable" and "intrinsically unscalable" subjects. However, this approach does not provide a solution for subjects who are scalable according to different models.

Mixture IRT models can be used to identify subgroups of subjects who differ in the use of the response scale. The origin of the differential use may vary from faking

and social desirability, to differential use or understanding of the middle category and the tendency of extreme responding. Also, influences of manifest variables, such as ethnic background, may result in qualitative differences in the response behavior. The advantage of mixture IRT models is that no assumptions as to the type of the qualitative differences have to be made beforehand. In this study, the latent classes will be characterized by their association with exogenous variables and there will be a closer look at the differences in the response behavior across latent classes. Furthermore, it will be examined whether the precision of the prediction can be improved by applying a mixture IRT model. In the next section, mixture IRT models are specified to describe the response behavior of a group of subjects. The estimated latent trait scores will be related to an external criterion measure in order to evaluate the improvement in prediction compared to the one class model.

4.2 Model Formulation

Both a mixture version of Masters' (1982) Partial Credit Model (mPCM) and Bock's (1972) Nominal Response Model (mNRM) will be analyzed. Let X_{ij} be a random variable denoting a response of a subject $i \in \{1, \ldots, I\}$ to an item $j \in \{1, \ldots, J\}$, with realizations $x_{ij} = k \in \{0, \ldots, K_j\}$. Let Y_i be a random variable denoting latent class membership with realizations $y_i = m \in \{1, \ldots, M\}$, and let θ_{im} denote the latent trait value of subject i in latent class m. For the mNRM, the probability that subject i from latent class m with latent trait value θ_{im} gives a response in category k of item j, is formulated as,

$$P(X_{ij} = k|\theta_{im}, Y_i = m) = \frac{\exp(\alpha_{jkm}\theta_{im} + \delta_{jkm})}{\sum_{h=0}^{K_j} \exp(\alpha_{jhm}\theta_{im} + \delta_{jhm})}, \tag{4.1}$$

with identifying restrictions $\sum_k \delta_{jkm} = 0$ and $\sum_k \alpha_{jkm} = 0$. The class-specific category difficulty and discrimination parameters are denoted by δ_{jkm} and α_{jkm} respectively. The mPCM is a special case of the mNRM, where the discrimination parameters are restricted to equal the category numbers. The mixture version of the PCM, based on the common formulation of the PCM, states the probability that a subject i gives a response in category

k of item j as given in Equation 4.2,

$$P(X_{ij} = k|\theta_{im}, Y_i = m) = \frac{\exp[\sum_{r=1}^{k}(\theta_{im} - \tau_{jrm})]}{1 + \sum_{h=1}^{K_j}\exp[\sum_{r=1}^{h}\exp(\theta_{im} - \tau_{jrm})]} \quad (4.2)$$

$$= \frac{\exp(k\theta_{im} + \delta_{jkm})}{\sum_{h=0}^{K_j}\exp(h\theta_{im} + \delta_{jhm})} \quad (4.3)$$

where τ_{jrm} denotes the threshold parameters, with $\tau_{j0m} = 0$. The threshold parameters can be reformulated as $\tau_{jrm} = \delta_{j,k-1,m} - \delta_{jkm}$, with the restriction $\delta_{j0m} = 0$. This re-parameterization allows reformulating the mPCM as given in Equation 4.3. The latent trait variable θ_{im} is transformed to θ_{im}/α_m, where the class-specific scaling parameter α_m describes the association between the latent trait variable θ_{im} and the items for latent class m. Transforming k to $\alpha_m k$ makes it clear that the discrimination parameters α_{jkm} of the mNRM are restricted in the mPCM to be equal to k. These values do not need to be estimated, but are equal to the category numbers k, which implies that the categories are ordered and spaced equidistantly (Heinen, 1996; Kelderman, 2007). The discrimination parameters of the mNRM can be considered class-specific item category scores, which are not fixed but need to be estimated. Consequently, the NRM does not require a priori ordering of the response categories. Particularly in personality measurement, with Likert scaling or the presence of categories like "?" and "Don't know", responses may be influenced by tendencies like preference for the middle category or extreme responding. To examine the functioning of the response scale, the mNRM will be fitted, where no a priori restrictions on the item parameters have to be made. To analyze whether a more parsimonious model fits the data, the mPCM is fitted as well. The models will be analyzed with different numbers of latent classes to identify the optimal number of classes to describe the data.

The exogenous variables ethnic background ($e \in \{1, \ldots, E\}$) and social desirability ($s \in \{1, \ldots, S\}$) are added to the model, as they both may influence the response behavior of the subjects. Smit, Kelderman, and Van der Flier (1999; 2000) have shown that by including exogenous variables, the measurement model becomes more stable. Standard errors of the parameter estimates as well as latent class assignment can benefit substantially. Furthermore, the exogenous variables may aid in the characterization of the latent classes (Smit, et al., 2003). Both latent variables are specified conditional on the exogenous variables. Note that ethnic background and social desirability are considered

fixed in all models, and are not influenced by the mixture solution. A saturated log-linear model was formulated for the association between the latent class variable and the exogenous variables;

$$\log P(Y_i = m|E_i = e, S_i = s) = \mu_{es} + \lambda_m + \lambda_{me} + \lambda_{ms} + \lambda_{mes}, \tag{4.4}$$

where μ_{es} is log the proportionality constant, and the λ parameters sum to zero over each index. The main effect of the latent class variable is described by λ_m, whereas the interaction effect of the latent class variable with ethnic background, social desirability, and both, are given by λ_{me}, λ_{ms}, and λ_{mes} respectively.

A linear regression model is formulated for the latent trait variable on the exogenous variables;

$$\theta_{im} = \beta_0 + \beta_e + \beta_s + \epsilon, \tag{4.5}$$

where $\epsilon \sim N(0, 1)$. The intercept is denoted by β_0, while β_e and β_s are the effects on the latent trait of ethnic background and social desirability respectively. The last category of both exogenous variables is specified as the reference category. Maximum likelihood estimates of the parameters of the simultaneous model, Equation 4.1 through 4.5, are computed by means of the EM-algorithm (Dempster, Laird, & Rubin, 1977). The models are analyzed with the program ℓEM (Vermunt, 1997). A graphical display of the model is given in Figure 4.1. The arrows denote a regression, while the squares and ellipses represent manifest and latent variables respectively.

For diagnostic purposes, individual latent trait estimates can be obtained once the best fitting mixture IRT model, among those compared, has been identified. Class-specific Maximum A Posteriori (MAP) estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iM})$ are obtained using

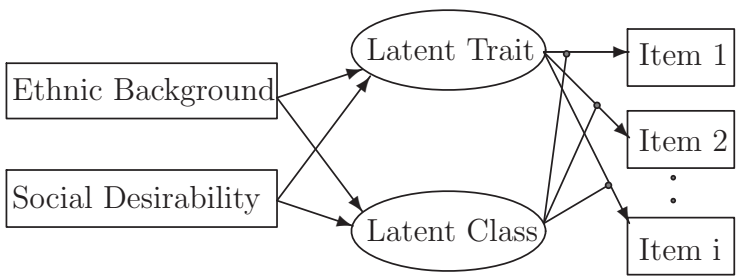

Figure 4.1: Discrete recursive graphical model with variables ethnic background, social desirability, latent trait, latent class and items 1 through i

the estimated item parameters with the program Multilog (Thissen, Chen, & Bock, 2003). Each person is assigned to latent class m with the highest posterior probability given his response pattern. The corresponding latent trait estimate of that particular latent class will be assigned to the person. For the standard measurement model, the one class version of the model is used. From observed response patterns and estimated item parameters, MAP-estimates of the latent trait are obtained.

The mixture IRT model gives a more detailed description of subjects response patterns, than the one class version of the model. Therefore, it is expected that the latent trait scores estimated using a mixture IRT model provides a better prediction of an external criterion, compared to the latent trait scores estimated with the one class model. The prediction will be evaluated for each latent class separately. It will be examined whether the precision of the prediction differs among latent classes.

4.3 Method

4.3.1 Data

The models described above will be applied to two scales of the Amsterdam Biographical Questionnaire (Amsterdamse Biografische Vragenlijst, or ABV; Wilde, 1970). The ABV is one of the most frequently used personality questionnaires in The Netherlands. It is based on the MPI (Maudsley Personality Inventory; Eysenck, 1959), which is an early version of the EPI (Eysenck Personality Inventory; Eysenck & Eysenck, 1975). The scales extraversion and neuroticism, with 21 and 30 items respectively, will be examined. Each item has three response categories, "yes", "?" and "no". For eight items of the extraversion scale, and six items of the neuroticism scale, two response categories were combined because they have equal scores in the original data.

The questionnaire also contains a scale measuring social desirability (in the EPI called Lie scale) consisting of 23 items, which is incorporated as an exogenous variable. As described earlier, social desirability is expected to influence the response behavior on personality scales. The social desirability scores are aggregated into three categories of equal proportion of subjects. The second exogenous variable is ethnic background, which has six levels, namely Dutch, Turkish, North-African, Antillean, Surinamese and Other.

A criterion measure is available for a part of the sample, corresponding to the scales

of extraversion and neuroticism. This criterion is a judgement about both extraversion and emotional stability by a psychologist based on a selection-interview. The psychologist had no knowledge of the test results of the applicant. The judgement was given on a five-point scale, of which the ends are respectively introvert (value of 1) and extravert (5), and emotional stable (1) and emotional instable (5).

4.3.2 Sample

The data were collected in two parts, each as an element of a personnel selection procedure for blue collar jobs and lower administrative jobs at the Dutch Railways, regional bus companies and road transport companies in The Netherlands. The first part of the data was previously analyzed by Smit, Kelderman, and Van der Flier (2003). The total sample, which contains all applicants in the two periods, consists of 3951 subjects. In Table 4.1 some demographic information about the sample is given. Eight subjects who had at least

Table 4.1: *Some Demographic Information About the Sample*

Ethnic background	N	Proportion men	Mean age
Dutch	812	0.87	28.24
Turkish	248	0.96	23.87
North-African	140	0.97	27.25
Antillean	112	0.85	30.71
Surinamese	475	0.83	30.29
Other	150	0.89	32.07
Unknown	2006	0.84	33.49

one missing value on one of the items, of the scales extraversion, neuroticism or social desirability, were left out of the analysis. The only missing data left are in the variable ethnic background and the criterion measure. Each subject showed missing data on one or both of these variables. The ethnic background and the criterion measure are both missing for 1662 subjects. For 344 subjects only their ethnic background is not recorded, while for 1937 only the criterion measure was missing. These missing data are missing by design, they were not recorded for a part of the sample. To handle the missing data, different types of observed frequency tables are constructed, belonging to the different subgroups of individuals for whom the same type of information is available. When estimating a model, all the information that is available in the different tables is used.

4.3.3 Analysis

First, both the mPCM and mNRM will be analyzed with different numbers of latent classes, for both of the scales extraversion and neuroticism. For mixture IRT models, a general problem is that local maxima may be found in addition to the global maximum likelihood solution one is looking for. Therefore, the models will be fitted ten times with random starting values (Rost, 1991). The best solution, defined by the lowest BIC-value, will be selected. The BIC (Schwarz, 1978) is preferred over the AIC as it takes the sample size into account, avoiding overparameterization (McLachlan & Peel, 2000). A choice will made between the two models, and the number of latent classes will be determined that provides the best fit to the data. In addition, some models with and without some further restrictions will be analyzed. The latent classes will be characterized on the basis of the item responses and exogenous variables. The estimated latent trait scores will be compared to the criterion data. The associations will be compared for each latent class separately.

4.4 Results

The fit statistics of the analyses of the extraversion (E) and neuroticism (N) scale are given in Table 4.2. The table shows for the mPCM and mNRM, with different numbers of latent classes ($\sharp$ m), the log-likelihood statistic ($\log \ell$), the number of parameters (npar), and the Bayesian Information Criterion (BIC). To compare models with different numbers of latent classes, the BIC-statistics were examined, where lower values indicate a better fit. For the mPCM of the extraversion scale, the model with five latent classes provided the best fit. For the mNRM, the model with three latent classes gave the best fit. Comparing the BIC-statistics of both models, the mNRM with three latent classes provided the best fit to the data. When examining the neuroticism scale, for both the mPCM and the mNRM the model with three latent classes provided the best fit. Comparing the BIC-statistics of both models, as for the extraversion scale, the mNRM with three latent classes provided the best fit to the data. The preference of the mNRM over the mPCM suggested that the assumption of equal distances between the response categories was too strong. The mNRM demonstrated that the response scale functions differently for different people, as has been shown previously (Hernández et al., 2004; Smit et al., 2003).

Table 4.2: *Fit Statistics for the Extraversion and Neuroticism Scale Modeled by the mPCM and mNRM With Different Numbers of Latent Classes*

Model	♯ m	log ℓ	npar	BIC
Extraversion				
mPCM	1	-61529.24	61	123563.54
	2	-60280.39	114	121504.64
	3	-59776.44	167	120935.59
	4	-59450.77	220	120723.08
	5	-59229.38	273	*120719.12*
	6	-59014.69	326	120728.56
mNRM	1	-60496.38	94	121771.04
	2	-59447.64	180	120385.63
	3	-59067.48	266	*120337.36*
	4	-58763.53	352	120441.52
Neuroticism				
mPCM	1	-67726.74	81	136124.13
	2	-65551.75	154	132378.58
	3	-64821.51	227	*131522.52*
	4	-64549.42	300	131582.74
mNRM	1	-66343.01	134	133795.50
	2	-64764.90	260	131682.52
	3	-64091.91	386	*131379.78*
	4	-63802.98	512	131845.17

In Table 4.3, the results are given of the analyses of the mNRM of both personality scales with three latent classes, and the same models with and without some further restrictions. Model 1 denotes the best fitting model found previously. Model 1a is the same model obtained by multiplying the main effects of the linear regression model, formulated for the latent trait and the exogenous variables as given in Equation 4.5, by a class-specific scaling parameters γ_m;

$$\theta_{im} = \beta_0 + \gamma_m^e \beta_e + \gamma_m^s \beta_s + \epsilon,$$

with additional identifying restrictions $\gamma_1^e = 1$ and $\gamma_1^s = 1$. This relaxes the assumption of class-invariance of the regression model.

Apart from the BIC-statistics, the likelihood ratio test could be used to compare the

Table 4.3: *Fit Statistics for the Extraversion and Neuroticism Scale Modeled by the mNRM With Three Latent Classes (Model 1), and the Same Model With/Without Some Further Restrictions*

Model	Effect	$\log \ell$	npar	BIC(log-L)
Extraversion				
1	mNRM Y=3	-59067.48	266	120337.36
1a	Model 1 with scaling parameter	-58999.15	270	120233.81
2	Model 1a without λ_{mes}	-59011.58	250	*120093.09*
3a	Model 2 without λ_{me}	-59062.96	246	120163.72
3b	Model 2 without λ_{ms}	-59145.20	240	120277.53
Neuroticism				
1	mNRM Y=3	-64091.91	386	131379.78
1a	Model 1 with scaling parameter	-64050.98	390	131331.04
2	Model 1a without λ_{mes}	-64065.54	370	*131194.57*
3	Model 2 without λ_{ms}	-64207.73	360	131396.12

relative fit of two nested models. The likelihood ratio test is defined by minus twice the differences in log-likelihood statistics. This approaches a chi-squared distribution, where the number of degrees of freedom equals the differences in number of free parameters of the models. For the extraversion scale, it is seen that Model 1a had a lower BIC-value than Model 1, indicating a better fit. The likelihood ratio test gave the same result, Δ -2 $\log \ell = 136.66$, Δ npar $= 4$, $p < .01$, preferring the less restricted model.

A saturated log-linear model was formulated for the association between the latent class variable and the exogenous variables, as shown in Equation 4.4. In Model 2, the three-way interaction term (λ_{mes}) was eliminated from Model 1a, to test whether all interaction effects were present. This improved the model fit even further, according to both the BIC and the likelihood ratio test, Δ -2 $\log \ell = 24.86$, Δ npar $= 20$, $p > .05$, preferring the more restricted Model 2. Finally, the presence of the two-way interaction terms λ_{me} (model 3a) and λ_{ms} (model 3b) was tested. The BIC-values of both models indicated a worse fit compared to Model 2. The same conclusion could be drawn from the likelihood ratio test, with Δ -2 $\log \ell = 102.76$, Δ npar $= 4$, $p < .01$, and Δ -2 $\log \ell = 267.24$, Δ npar $= 20$, $p < .01$ for Model 3a and 3b respectively.

For the neuroticism scale, adding the class-specific scaling parameter to Model 1 improved the fit. This was indicated by both a lower BIC-value and the likelihood

ratio test, Δ -2 $\log \ell = 81.86$, Δ npar $= 4$, $p < .01$. The model improved even further by eliminating the three-way interaction λ_{mes}. There was a lower BIC-value and the likelihood ratio test, Δ -2 $\log \ell = 29.12$, Δ npar $= 20$, $p > .05$, indicated that the more restricted model should be preferred. The model without the interaction term λ_{me} was analyzed several times with random starting values, unfortunately the model was not estimable. Removing the interaction term λ_{ms} gave a worse fit according to both the BIC and the likelihood ratio test, Δ -2 $\log \ell = 284.38$, Δ npar $= 10$, $p < .01$. In conclusion, Model 2 was also the best fitting model for the neuroticism scale.

It was examined whether the estimated item parameters of some of the items were actually invariant across latent classes. The equality of the estimated item parameters across the latent classes was examined by analyzing the best fitting model with equality restrictions on the parameters of one item across latent classes. This was performed for each item separately. The models with equality restrictions always fitted worse, based on both the BIC-statistics and likelihood ratio test, than the model where all item parameters vary across latent classes.

4.4.1 Associations with exogenous variables

There was a significant effect of ethnic background and social desirability on the latent class variable, for both personality scales. The relation between the latent classes and the exogenous variables for the extraversion scale, is given in Figure 4.2. It is seen that Dutch subjects had a high probability to belong to the first latent class, and a small probability to belong to latent class two. The subjects that give highly socially desirable answers had a higher probability of being in the second latent class. Subjects with a low tendency to give a socially desirable answer were more likely to be in the third latent class. In this latent class all ethnic groups were represented. These results were consistent with the interaction parameters of the latent class variable with each of the exogenous variables, λ_{me} and λ_{ms}, as shown in Table 4.4. Higher absolute values of λ_{me} and λ_{ms} indicate a stronger interaction between the latent class variable and both of the exogenous variables. It was shown that λ_{mes} could be removed from the model, which indicates that the effects of the exogenous variables on the latent class variable are independent.

Comparable results were found for the neuroticism scale, see Table 4.4 and Figure 4.3. Again, Dutch subjects were most likely to belong to the first latent class. Subjects

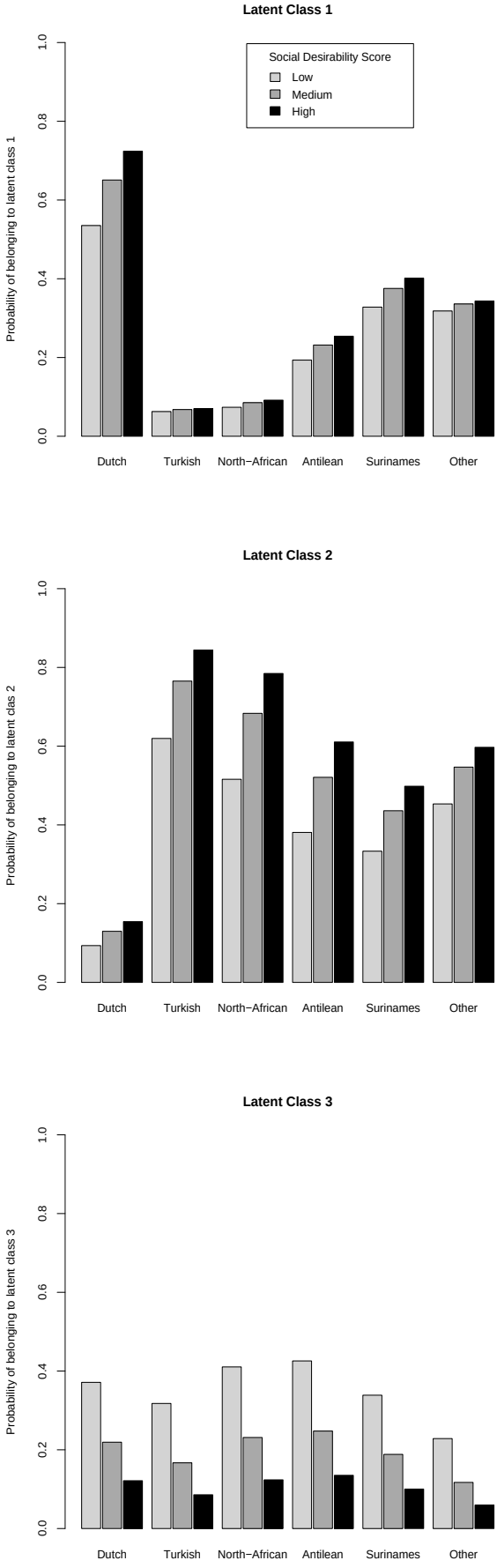

Figure 4.2: Probability of belonging to each latent class, given ethnic background and social desirability, for the extraversion scale

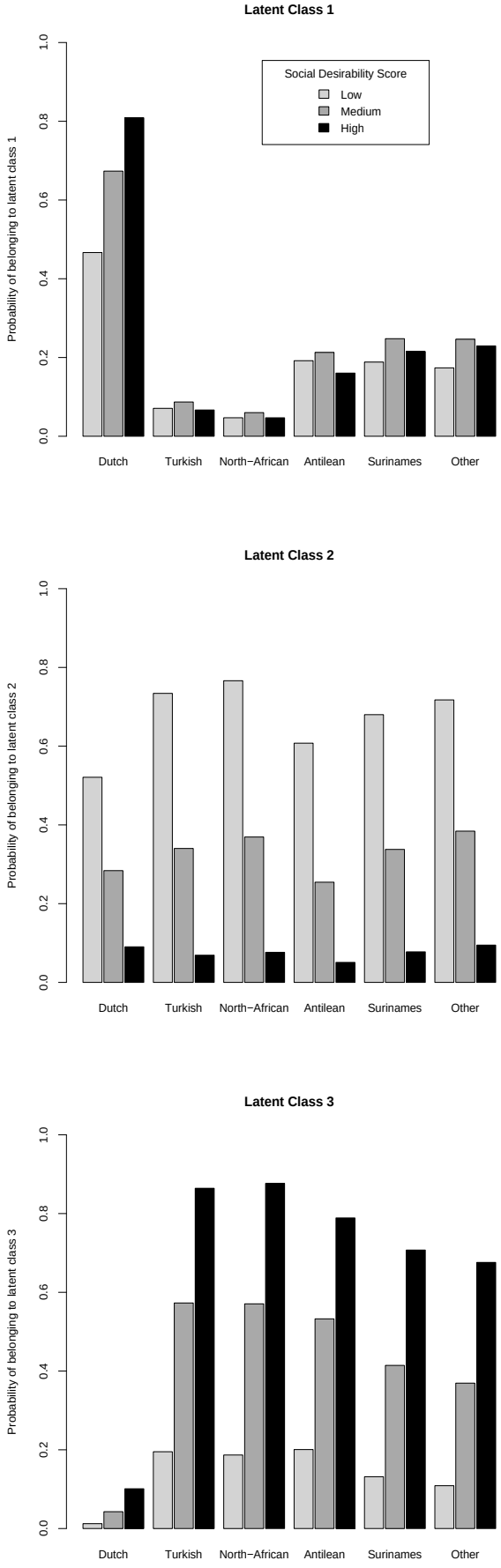

Figure 4.3: Probability of belonging to each latent class, given ethnic background and social desirability, for the neuroticism scale

Table 4.4: *Interaction Parameters of the Latent Class Variable With Ethnic Background (λ_{me}) and Social Desirability (λ_{ms}), for the Extraversion and Neuroticism Scale*

	Extraversion			Neuroticism		
	Latent class			Latent class		
Category	1	2	3	1	2	3
Ethnic background						
Dutch	1.0929	-1.2424	0.1495	1.5661	0.1606	-1.7267
Turkish	-0.9151	0.7852	0.1299	-0.7237	0.0985	0.6253
North-African	-0.8329	0.5245	0.3084	-0.9975	0.2784	0.7191
Antillean	-0.0995	-0.0121	0.1116	-0.0059	-0.3690	0.3749
Surinamese	0.3731	-0.2008	-0.1723	0.0846	-0.1473	0.0627
Other	0.3815	0.1454	-0.5270	0.0764	-0.0212	-0.0552
Social desirability						
Low	-0.2010	-0.3118	0.5128	-0.0951	0.9973	-0.9022
Medium	-0.0047	0.0175	-0.0128	-0.0613	0.0571	0.0042
High	0.2057	0.2943	-0.5000	0.1563	-1.0543	0.8980

who have a low tendency to give socially desirable responses were most likely to belong to the second latent class. This resembled latent class three of the extraversion scale. Subjects with a higher tendency to give a social desirable answer were likely to belong to latent class three. Also, there were hardly any Dutch subjects in this latent class.

4.4.2 Response tendencies

The way the subjects used the response scale, differed among latent classes. Examining the Category Response Functions (CRF's) gives insight into the preference for one or more item categories in a particular latent class. The solid line is the indicative category, the dashed line is used for the middle category, and the dotted line for the contra indicative category. Note that the response categories, "yes", "?" and "no", as shown in de legend, do not always follow the same order. For both personality scales, two items will be discussed that point out the differences among the latent classes. Higher levels of theta indicate a higher degree of extraversion, as opposed to introversion, and more neuroticism, as opposed to emotional stability.

The CRF's of Item 1: "Do you prefer to keep your social life limited to a few

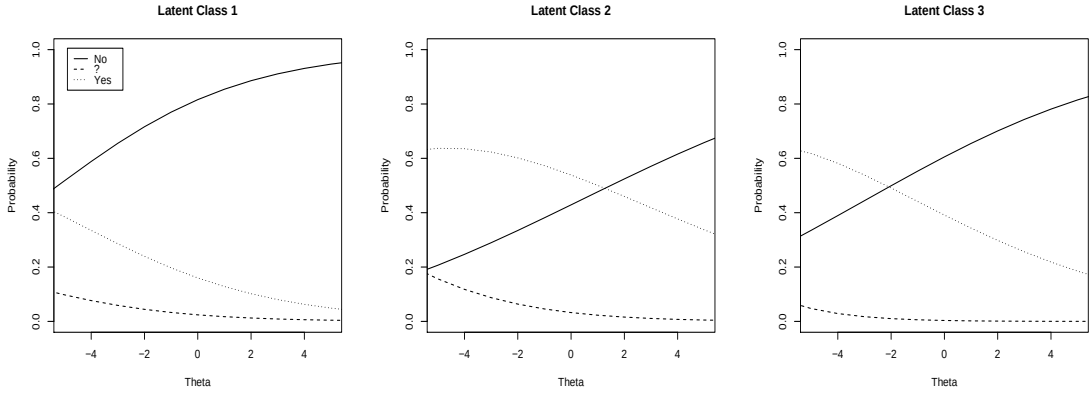

Figure 4.4: Category response functions for extraversion Item 1

good friends ?" of the extraversion scale, are shown in Figure 4.4. It is seen that the "?" category was seldom used in all latent classes. In the first latent class, the response "no" seemed to be preferred across all levels of the latent trait. In the second latent class, where all ethnic minority groups were represented, the response "yes" was given relatively often. Te Nijenhuis, Van der Flier and Van Leeuwen (1997) found that immigrants tended to respond "yes" to this item. They argued that in particular Turks and North Africans prefer to keep their social intercourse limited to their in-group. Subjects from the third latent class responded "yes" too, but were more inclined to respond "no" as well.

In general, subjects from the third latent class seemed to avoid the "?" category. This was shown for item 1, but can also be seen for, for example, item 9: "Do you consider yourself to be a communicative person ?", as shown in Figure 4.5. It is seen in Table 4.4 that the third latent class could be characterized by lower social desirability scores. As described earlier, extraversion is seen as desirable, which suggests that "yes"

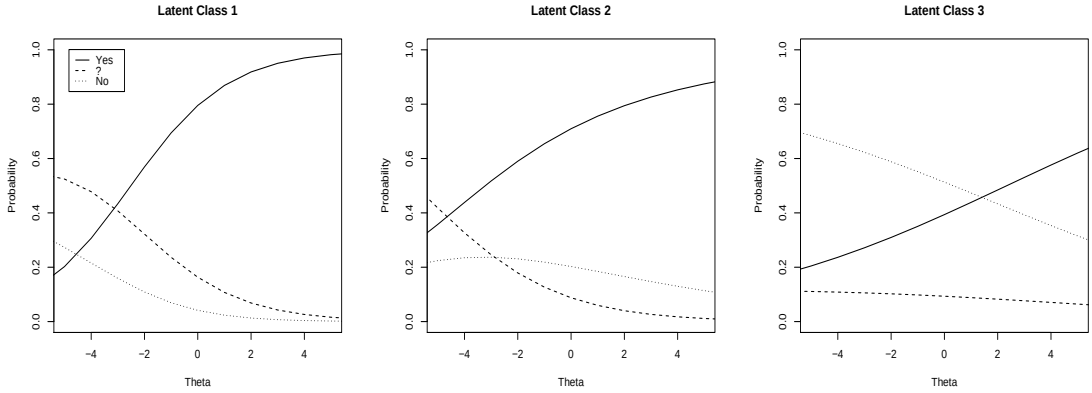

Figure 4.5: Category response functions for extraversion Item 9

is the desirable answer to this item. As expected the subjects from latent class three responded "yes" less often compared to the other latent classes. Subjects from the first and second latent class may respond in the "?" category, when they have lower latent trait scores.

It has been shown that the response scale functioned differently across the latent classes. Subjects from the first and second latent class seemed to use all response categories, including the "?" category, whereas this category was avoided by subjects from latent class three. Furthermore, these subjects felt no need to give socially desirable responses, which may be a reason for little responding in the "?" category. The results from Table 4.4 and the CRF's given above, show that subjects from the first and second latent class gave more socially desirable responses.

Next, the neuroticism scale. Item 14: "Is it true, that sometimes you feel full of energy and activity, sometimes you feel very indolent and lifeless ?", was representative for many items, see Figure 4.6. It is seen for latent class one, that subjects scoring higher

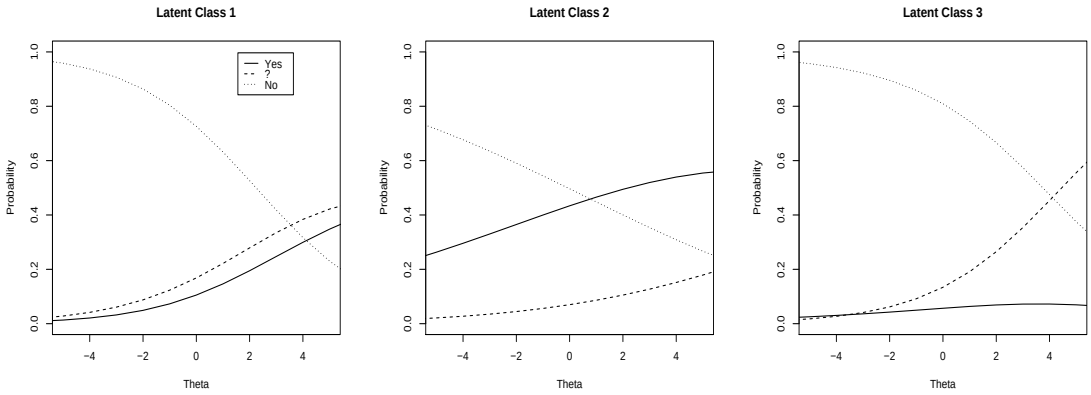

Figure 4.6: Category response functions for neuroticism Item 14

on the latent trait seemed to doubt between the "?" and "yes" category, often preferring the "?" response. In the second latent class, the subjects tended to give relatively often a less socially desirable response, and they avoided the "?" category. Subjects from the third latent class seemed to prefer the "?" as they scored higher on the latent trait, and seemed to avoid a "yes" response.

Item 21 stated: "Are you often moody ?", see Figure 4.7. Subjects in latent class one responded "no". In the second latent class, higher scoring subjects responded "yes" relatively often, which was again, the socially undesirable response. The "?" response

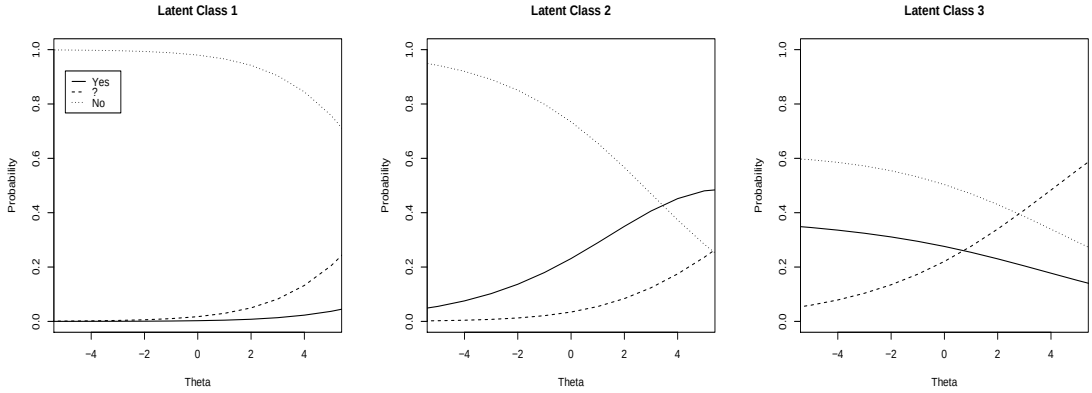

Figure 4.7: Category response functions for neuroticism Item 21

was given in the third latent class by higher scoring subjects. It is remarkable, however, that the "yes" response was given relatively often by lower scoring subjects in this latent class. The pattern could be explained by the wording of the item, in combination with the presence of relatively many subjects from ethnic minority groups. These subjects probably misunderstood the question, since it could be mistaken for "Are you often in a good mood ?".

Overall, for the neuroticism scale two response tendencies could be shown. The use of the "?" response category differed among the latent classes. Subjects from latent class two avoided this category, whereas in the third latent class it was used relatively often. In the first latent class the subjects did not seem to prefer or avoid any category. The second response tendency concerned social desirability. For the second latent class this seemed to be no issue, these subjects seemed to respond honestly, maybe therefore not using the "?" category. Subjects from the third latent class, with higher social desirability scores, seemed to hesitate in giving undesirable responses, preferring the "?" category.

4.5 Prediction

The subjects for whom criterion data were available were assigned to a latent class, based on the conditional probability of latent class membership given their response pattern. Based on their response patterns the corresponding latent trait score was estimated for that particular latent class. The number of subjects assigned to each of the latent classes for both personality scales, as well as their proportions, are given in Table 4.5. The proportions of subjects of the total sample within each latent class for whom criterion

Table 4.5: *Number and Proportion of Subjects (for Whom Criterion Data Are Available) Assigned to Each Latent Class for the Extraversion and Neuroticism Scale*

Latent class	Extraversion		Neuroticism	
	N	Proportion	N	Proportion
1	164	0.48	174	0.51
2	109	0.32	106	0.31
3	71	0.20	62	0.18
Total	344	1.00	342	1.00

data were available were 0.089, 0.083 and 0.091 respectively for the extraversion scale. For the neuroticism scale, these proportions were 0.106, 0.083 and 0.061.

To evaluate the improvement of the prediction of an external criterion measure, the estimated latent trait scores were related to the criterion measure. This was done for each latent class separately. The results for the extraversion scale are shown in Table 4.6. It is seen that the latent trait scores estimated with a standard measurement model,

Table 4.6: *Class-specific Polyserial Correlations for the Relation of the Criterion Measure With the Latent Trait Estimates of the One Class Model and the Mixture IRT Model for the Extraversion Scale*

	Latent class 1	Latent class 2	Latent class 3
One class model	.193	.318	.375
Mixture IRT model	.084	.247	.226

the one class model, provided more accurate estimates of the criterion measure than the mixture IRT model. This result appeared for each latent class. The differences between the polyserial correlations (Cox, 1974) were significant for latent class 1 and 3, but not for latent class 2 ($z = 3.61, p < .01; z = 1.54, p = .06; z = 1.97, p = .025$, for the three latent classes respectively). The mixture IRT model may describe the response behavior of the subjects in more detail, but the latent trait estimates did not provide an improvement in the prediction of the criterion. The Pearson correlations between the estimated latent trait scores of the one class model and the mixture IRT model within each latent class are .925, .876, and .785 for the three latent classes respectively.

For the neuroticism scale, the results were more promising. It is seen in Table 4.7

Table 4.7: *Class-specific Polyserial Correlations for the Relation of the Criterion Measure With the Latent Trait Estimates of the One Class Model and the Mixture IRT Model for the Neuroticism scale*

	Latent class 1	Latent class 2	Latent class 3
One class model	.274	.227	.416
Mixture IRT model	.302	.274	.577

that the correlations of the latent trait estimates for the mixture IRT model with the criterion were higher than for the one class model. These differences were significant for latent class 1 ($z = -1.91, p = .03$) and latent class 3 ($z = -2.81, p < .01$), but not for latent class 2 ($z = -1.37, p = .09$). The mixture IRT model improved the prediction of the criterion for the neuroticism scale. The class-specific Pearson correlations between the estimated latent trait scores of the one class model and the mixture IRT model are .980, .935, and .867 respectively.

4.6 Discussion

Mixture IRT models proved to yield a better fit to the data than a one class model, for the extraversion and neuroticism scale of the ABV. In the study by Smit, Kelderman, and Van der Flier (2003) no more than two latent classes were allowed. Based on a larger data set it was shown that this is too restrictive, and a larger number of latent classes should be considered. The mNRM was preferred to the mPCM. Although restrictive measurement models have the advantage of simplicity, a more lenient model may provide a better solution. These models leave more room to detect qualitative differences in the response patterns.

The mNRM with three latent classes was selected as the best fitting model. The latent classes differed with respect to social desirability, ethnic background and the use of the response scale. Other variables, like gender or age, could also have been relevant exogenous variables. However, it was decided to focus on ethnic background and social desirability, as these have been shown to influence the response behavior of subjects. For both personality scales, one class with lower social desirability scores was identified, and two latent classes with a higher tendency to give a socially desirable response. These two latent classes differed with regard to ethnic background, one latent class with relatively

many Dutch subjects, and one class with subjects from ethnic minority groups. Lower social desirability seemed to be related to less use of the "?" category. These differences were more pronounced for the neuroticism scale than for the extraversion scale.

The characterizations of the latent classes were quite comparable across the two scales; of course the order of the classes is arbitrary. Application of mixture IRT models improved the prediction for the neuroticism scale, but not for the extraversion scale. These results were consistent across latent classes. In personality measurement, like in the present study, correlations with external criteria are usually not very high. Moreover, it has been shown that correlations between self-ratings and ratings by others, such as a psychologist, are lower than self-self or other-other ratings (Kolk, Born, & Van der Flier, 2004). Overall, it is concluded that mixture IRT models provide possibilities to improve the prediction of external criteria, but that it may vary across personality scales.

An important remaining issue is the comparability of the latent trait scores across latent classes. As the measurement models differ across classes, the latent trait scores cannot be compared directly. Only after latent trait estimates within different latent classes have been transformed to be on the same scale, comparison of subjects across latent classes can be possible. A method commonly used to link IRT scales, is to impose equality constraints on item parameters across latent classes for a subset of items (e.g., Kelderman & Macready, 1990; Kolen & Brennan, 1995; Von Davier & Yamamoto, 2004). However, this requires strong assumptions about which items will function the same way across latent classes. Assuming that the same construct is measured in all latent classes, an alternative would be to transform the latent trait estimates to be on a common scale based on their associations with an external criterion. If the resulting scores show a better prediction than the scores based on the one class model, this procedure is justifiable from a practical point of view. However, it should be noted that the common scale is criterion specific, depending on the value of the determination coefficient.

A justification of the assumption of a common construct in each latent class can be found in a non-significant test of the invariance of a subset of item parameters, after they have been equated, across latent classes. The converse is not necessarily true, if all test items function differentially, it does not necessarily mean that the construct differs across latent classes. Another justification may be found in the latent traits' empirical relations with other variables in the nomological network (Cronbach & Meehl, 1955). If

the latent traits' relations with those variables are invariant across latent classes, we may safely assume that the equated trait scores measure the same construct.

Another, following important issue concerns the use of latent class information. In this study, the subjects were assigned to a latent class, and obtained the corresponding latent trait estimate. However, latent class probabilities differ among subjects. Some subjects have a high probability of belonging to a certain latent class, while for others the probabilities are more or less equally distributed over the latent classes. The certainty with which the subjects are assigned to the latent classes is not taken into account, but may influence the accuracy of the prediction.

This study demonstrates the possible advantages of applying mixture IRT models to personality questionnaires. When heterogeneity in the population, with regard to response tendencies or other manifest variables, can be expected, mixture IRT models can help to gain more insight in the response behavior of the subjects and provide possibilities of improving the prediction of external criteria. An important advantage of mixture IRT models, compared to DIF analyses, is that no a priori grouping assumptions have to be made. If the differences in response behavior are limited to a small number of items, mixture IRT models could also be used in test construction (e.g. Kelderman & Macready, 1990).

Chapter 5

The Use of Latent Class Membership Probabilities in Latent Trait Estimation and Prediction on the Basis of Mixture Item Response Theory Models

In mixture IRT modeling, subjects are usually assigned the latent trait estimate corresponding to the latent class with the highest probability. The certainty of latent class assignment is not taken into account. An alternative to assignment is to weigh the latent trait estimates with their corresponding latent class probabilities. A simulation study is conducted, showing the influence of differences in item parameters across latent classes, on the performance of weighted and assigned latent trait estimates. Weighted latent trait estimates showed equal or higher correlations with a simulated criterion and the simulated latent trait values compared to assigned latent trait estimates. Data from two personality scales show that weighted latent trait estimates may provide an equally good or better prediction of an external criterion especially for shorter item sets, compared to assigned latent trait estimates.

Mixture IRT models allow for qualitative as well as quantitative differences in the response process of subjects. Subgroups of subjects are identified that respond according to different measurement models, or the same model but with different parameter estimates

This chapter has been submitted for publication.

across the latent classes (Rost, 1990; 1991). Often, as is common in bias research, between-group differences are based on a partitioning of a sample of subjects on the basis of manifest variables such as gender, race or age. Mixture IRT models make no assumption as to the type or cause of the qualitative differences in the responses beforehand. Qualitative differences may be related to the variables mentioned, but may also be due to response tendencies, such as social desirability, developmental stages or other factors. Exogenous variables may be added to the mixture IRT model, to examine the nature of the differences between the latent classes.

Mixture IRT models have been studied from a variety of perspectives. Kelderman and Macready (1990) presented a log-linear Rasch model with a latent grouping variable to assess differential item functioning across latent classes (see also Kelderman, 2007). A mixture version of the linear logistic test model was analyzed by Mislevy and Verhelst (1990) to assign subjects to classes that correspond to item-solving strategies. Strategy shifts in problem solving have been studied with a mixture Rasch model (Rijkes & Kelderman, 2007). The saltus model (Wilson, 1989) and latent class models (Jansen & Van der Maas, 1997) have been used to analyze qualitative differences related to developmental stages. Mixture versions of the nominal response model and the generalized partial credit model were studied by Smit, Kelderman and Van der Flier (2003). Their analyses revealed qualitative differences between subjects in their response process, indicating that the parameters of the "?" response category were not invariant over subjects (see also Hernández, Drasgow, & González-Romá, 2004). Zickar, Gibby, and Robie (2004) applied the mixture Rasch model to uncover faking subgroups among a sample of applicants and incumbents. Finally, Maij-de Meij, Kelderman and Van der Flier (in press) showed that mixture IRT models provide possibilities for improvement in the prediction of external criteria. Von Davier and Carstensen (2007) provide more examples of studies using mixture IRT models.

Usually, latent class membership probabilities are used for the assignment of subjects to a latent class. A subject i is assigned to latent class $m \in \{1, \ldots, M\}$ with the highest probability $\hat{\pi}_{mi}$ of latent class membership given the item response pattern. However, the differences in latent class probabilities across latent classes differ among subjects. Some subjects have a high probability of belonging to a certain latent class, which offers a high probability of correct allocation. For other subjects the probabilities of latent class

membership do not differ much, providing less certainty that they are correctly allocated. This may be of importance when estimating a latent trait score for subjects.

Subjects are usually assigned the latent trait estimate $\hat{\theta}_{mi}$ of the latent class m to which they have been assigned. If the same construct is measured in each latent class and it is measured on the same scale, an alternative to assignment is to compute a weighted latent trait estimate. The individuals latent trait estimates are weighted by their corresponding latent class probabilities:

$$\hat{\theta}_i = \sum_{m=1}^{M} w_{mi}\, \hat{\theta}_{mi}, \tag{5.1}$$

where the class-specific weights w_{mi} are equal to the latent class probabilities $\hat{\pi}_{mi}$. Assignment is a special case of weighting, where the weights w_{mi} are equal to 1 or 0, for the highest and remaining latent class probabilities $\hat{\pi}_{mi}$ respectively. The assignment and weighting procedures will result in nearly the same latent trait estimates, if a subject is assigned with great certainty to a latent class. The differences may be more pronounced, however, if a subject has about equal probability of belonging to different latent classes. The differences between weighting and assignment should increase as the variances of the individual latent class probabilities, $\text{var}_m(\hat{\pi}_{mi})$ decrease.

The $\text{var}_m(\hat{\pi}_{mi})$ is influenced by the differences in the item parameters across latent classes. When the item parameters across latent classes become more similar, a response pattern becomes less typical for one particular latent class, resulting in lower $\text{var}_m(\hat{\pi}_{mi})$. As the differences in the item parameters across latent classes become larger, a response pattern may fit a specific IRT model better compared to the models of the other latent classes. This should result in increasing $\text{var}_m(\hat{\pi}_{mi})$. In short, as the item parameters vary increasingly across latent classes, $\text{var}_m(\hat{\pi}_{mi})$ should increase, which in turn should decrease the differences between weighting and assignment.

The differences in the item parameters across latent classes also influence the differences in the latent trait estimates $\hat{\theta}_{mi}$ across latent classes. As the item parameters vary increasingly across latent classes, these differences become larger, which should result in larger differences between weighted and assigned latent trait estimates. This means that there are two competing effects, related to weighting and assignment of latent trait estimates in mixture IRT modeling. As the differences in the item parameters become larger across latent classes, $\text{var}_m(\hat{\pi}_{mi})$ becomes larger, decreasing the differences between

weighted and assigned latent trait estimates. On the other hand, the differences between weighting and assignment should increase in consequence of larger differences in $\hat{\theta}_{mi}$. So far, no study has examined a weighted latent trait estimate in mixture IRT modeling, as an alternative to assigning a latent trait estimate. It is not clear whether the differences between weighting and assignment will increase or decrease as the item parameters vary increasingly across latent classes. Furthermore, test length may influence the precision of prediction. Longer tests may provide both more accurate latent trait estimates, and increase $\text{var}_m(\hat{\pi}_{mi})$, which in turn affect the differences between weighting and assignment.

In this study, it will be examined whether a weighted latent trait estimate provides a more accurate prediction of an external criterion and a more accurate estimate of the latent trait, compared to an assigned latent trait estimate. First, a simulation study is conducted, focussing on the specific effects of differences in item difficulty and discrimination parameters across latent classes, on the prediction of simulated criterion and latent trait values by assigned and weighted latent trait estimates. It will be examined whether the precision of prediction is influenced by $\text{var}_m(\hat{\pi}_{mi})$ and/or the differences in $\hat{\theta}_{mi}$. This will be studied for a short and longer test. The precision of prediction is evaluated by examining the correlations of the estimated latent trait values with the true latent trait values. Because an empirical example is analyzed as well, which naturally does not contain true latent trait values, also the correlations with a related criterion measure will be evaluated. Next, the empirical example will be given, to evaluate the differences between assignment and weighting for a real data sample. For both the simulation study and the empirical example, the results will also be compared to the performance of a one-class model in predicting criterion and/or latent trait values.

5.1 Method

In mixture IRT modeling, a measurement model is formulated where (some of) the item parameters may vary across latent classes. Let X_{ij} denote a response of a subject i to an item $j \in \{1, \ldots, J\}$, with realizations $x_{ij} = k \in \{0, \ldots, K_j\}$. Let Y_i be a random variable denoting latent class membership of subject i with realizations $y_i = m \in \{1, \ldots, M_i\}$, and let θ_{mi} denote the latent trait value of a subject in latent class m. For the mixture version of Bock's (1972) nominal response model (Smit et al., 2003; Maij-de Meij, Kelderman, &

Van der Flier, in press), the probability that a subject i from latent class m with latent trait value θ_{mi} gives a response in category k of item j, is formulated as,

$$P(X_{ij} = k|\theta_{mi}, Y_i = m) = \frac{\exp(\alpha_{jkm}\theta_{mi} + \delta_{jkm})}{\sum_{h=1}^{K_j} \exp(\alpha_{jhm}\theta_{mi} + \delta_{jhm})}, \tag{5.2}$$

with identifying restrictions $\sum_k \delta_{jkm} = 0$ and $\sum_k \alpha_{jkm} = 0$ on the class-specific category difficulty and discrimination parameters respectively. Assuming normality of the latent trait, maximum likelihood estimates of the model parameters can be computed by means of the EM-algorithm (Dempster, Laird, & Rubin, 1977; Vermunt, Ver97). Mixture IRT models are fitted with different numbers of latent classes to identify the optimal number of latent classes to describe the data. Manifest, exogenous variables may be added to the model, as indicators of the latent class variable. It has been shown that by including exogenous variables, the measurement model becomes more stable (Smit, Kelderman, & Van der Flier, 1999; 2000). Standard errors of the parameter estimates as well as latent class assignment can benefit substantially. Furthermore, the exogenous variables may aid in the characterization of the latent classes (Smit et al., 2003).

Once the best fitting model, among those compared, is identified, the individual latent trait estimates and latent class probabilities can be obtained. The posterior latent class probabilities for each subject, $\hat{\pi}_{mi}$, are obtained from the joint response model for $\boldsymbol{X}_i$=$(X_{i1}, \ldots, X_{iJ})$, the prior distribution, $f(\theta)$, for the latent trait, and $P(Y = m) = \pi_m$. Using Bayes formula (Box & Tiao, 1992) we have

$$\begin{aligned} \hat{\pi}_{mi} = P(Y = m|\boldsymbol{X}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}) &= [P(\boldsymbol{X}_i|Y = m, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})P(Y = m)]/P(\boldsymbol{X}_i) \\ &= [P(Y = m) \int P(\boldsymbol{X}_i|Y = m, \theta, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})f(\theta)d\theta]/P(\boldsymbol{X}_i), \end{aligned} \tag{5.3}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\delta}}$ are the estimated item parameters. Assigning the subject to the latent class with the largest posterior probability is called the maximum a posteriori (MAP) estimate of Y. Similarly, class-specific MAP estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iM})$ of the latent trait can be obtained.

It is assumed that the same construct is measured within each of the latent classes. Still, the latent trait need not to be on the same scale across latent classes. This is required, however, for the weighting procedure as it combines class-specific latent trait estimates. Also, using assigned latent trait estimates across latent classes demands that the trait estimates are on the same scale.

The focus is on prediction. Therefore, the class-specific latent trait estimates are transformed to be on the scale of an external criterion measure, where the criterion is denoted by Z_i with realizations $z_i = r \in \{1, \ldots, R_i\}$. A linear regression model is specified for each latent class separately,

$$z_i = \beta_{0m} + \beta_{1m}\hat{\theta}_{mi} + \epsilon_m, \tag{5.4}$$

where $\epsilon_m \sim N(0, \sigma^2)$. As the subjects differ in their latent class membership probabilities and their latent class membership is unknown, two regression approaches can be followed to transform the latent trait variables to be on a common scale:

Scale transformation Method I Each subject is assigned to the latent class with the highest posterior probability, $\max_m \hat{\pi}_{mi}$, and obtains the corresponding latent trait estimate. The assigned latent trait estimates for these particular latent classes, and the corresponding criterion data, are used to estimate the class-specific regression parameters. Since only the assigned latent trait estimates are used, the number of subjects of whom the data are used to estimate the regression parameters differs for each latent class.

Scale transformation Method II For each latent class, the latent class probabilities, $\hat{\pi}_{mi}$, serve as weights when regressing the criterion on the latent trait estimates. The regression is formulated for each latent class separately. This method uses all information available, so for each latent class, data of all subjects are used to estimate the regression parameters.

The regression parameters are estimated using the criterion data, and the final latent trait estimates may be related to the criterion as well. Therefore, to compare correlations with the criterion measure, cross-validation is needed. The sample of the subjects for whom criterion data are available is randomly split in two subsamples. Both subsamples are used separately to estimate class-specific regression parameters. Subsequently, these regression parameters are used to transform the class-specific latent trait estimates of the other subsample. The transformation assures that for each subject the latent trait estimates can be compared with those of other subjects, both within and across latent classes. Next, each subject has two latent trait estimates for each latent class, one transformed by the regression parameters of scale transformation Method I, and one transformed by the

parameters of Method II. Next, the assignment and weighting procedures, described in the Introduction, are followed to obtain a weighted and an assigned latent trait estimate for both methods of scale transformation. To evaluate which of the four combinations of the procedures is the most accurate, the final latent trait estimates are related to the external criterion measure and/or true latent trait values.

Furthermore, a one-class model is analyzed. The maximum likelihood estimates of the item parameters are subsequently used to estimate latent trait values, assuming this one-class model would fit the data. Correlations with the (simulated) criterion and/or latent trait values are examined, to evaluate whether the mixture IRT model improves the prediction compared to a one-class model.

5.2 Simulation Study

A simulation study is performed, to determine whether a weighted latent trait estimate is a better alternative for prediction compared to an assigned latent trait estimate. As described in the Introduction, two effects are associated with differences in weighted and assigned latent trait estimates. The latent trait estimates are affected by the differences in $\hat{\theta}_{mi}$ as well as by $\text{var}_m(\hat{\pi}_{mi})$. In this simulation study, it will be examined what the trade-off is between these two competing effects. Also, the influences of the item parameters are investigated.

5.2.1 Design

Data sets of five thousand subjects are generated according to a mixture version of the Birnbaum model, with two latent classes of equal size. The Birnbaum model is equivalent to a nominal response model (Equation 5.2) for dichotomous responses, where the item discrimination parameters (α_{jm}) and/or difficulty parameters (δ_{jm}) may vary across latent classes m;

$$P(X_{ij} = 1|\theta_{mi}, Y_i = m) = \frac{\exp(\alpha_{jm}(\theta_{mi} - \delta_{jm}))}{1 + \exp(\alpha_{jm}(\theta_{mi} - \delta_{jm}))}.$$

The latent trait and a criterion are both sampled from a multivariate standard normal distribution, with a correlation of .45.

The design of the simulation study is a 2 x 2 x 3 design, with 9 or 18 items, where the difficulty parameters may vary a little or a lot, and the discrimination parameters do

Table 5.1: *Design of the Simulation Study*

Number of Items	Difficulty	Discrimination		
		Equal	Small Inequality	Large Inequality
9	Small Inequality	a1	a2	a3
9	Large Inequality	b1	b2	b3
18	Small Inequality	A1	A2	A3
18	Large Inequality	B1	B2	B3

not vary, vary a little or vary a lot across latent classes. An overview of the design of the simulation study is given in Table 5.1. The parameter sets used in this simulation study for Conditions a1 through b3, are given in Table 5.2.

For Conditions A1 through B3 the parameters are equal to those of a1 through b3, with twice the number of items, where the parameter values of Item 10 through 18 are equal to the parameter values of Item 1 through 9. The difficulty parameters are in the range of -1 through 1, which, with nine items and equal distances, results in increases of 0.25 for each subsequent item. For the second latent class of Condition a, with small inequalities across latent classes, the difficulty parameters of Item 1 and 3 are exchanged, as well as the parameters of Item 7 and 9. Larger inequality (Condition b) is obtained by exchanging de difficulty parameters of Item 2 and 4 and of the Items 6 and 8 as well for the second latent class. For the discrimination parameters, a simple design is chosen

Table 5.2: *Difficulty (Conditions a and b) and Discrimination (Conditions 1 through 3) Parameter Sets of the Simulation Study, Conditions a1 through b3*

Condition	Class	Item								
		1	2	3	4	5	6	7	8	9
a	1	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0
	2	-0.5	-0.75	-1.0	-0.25	0.0	0.25	1.0	0.75	0.5
b	1	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0
	2	-0.5	-0.25	-1.0	-0.75	0.0	0.75	1.0	0.25	0.5
1	1	0.7	1.3	0.7	1.3	1.0	1.3	0.7	1.3	0.7
	2	0.7	1.3	0.7	1.3	1.0	1.3	0.7	1.3	0.7
2	1	0.7	1.3	0.7	1.3	1.0	1.3	0.7	1.3	0.7
	2	1.3	0.7	1.3	0.7	1.0	0.7	1.3	0.7	1.3
3	1	0.4	1.6	0.4	1.6	1.0	1.6	0.4	1.6	0.4
	2	1.6	0.4	1.6	0.4	1.0	0.4	1.6	0.4	1.6

with three different values balanced across the items. The discrimination parameter of Item 5 is set equal to 1, for the other items the values are 0.7 and 1.3. The discrimination parameters for the second latent class of the small inequality condition are all (but Item 5) exchanged. For the large inequality condition, the differences across latent classes are increased by doubling the differences in the parameter values of Condition 2, from 0.6 to 1.2, and keeping 1.0 as the mean value. For each condition, five hundred data sets are generated.

The item parameters, latent trait distribution and latent class sizes are known, and used to compute the individual class-specific MAP estimates of the latent trait $\hat{\theta}_{mi}$, and the posterior latent class probabilities $\hat{\pi}_{mi}$. The class-specific latent trait estimates do not need a scale transformation, as they are on a common scale by design. Next, weighted and assigned latent trait estimates are computed as described in the Introduction. These estimated latent trait values are correlated with the simulated criterion and latent trait values. First, it will be examined what the influence of differences in item parameters across latent classes is on $\text{var}_m(\hat{\pi}_{mi})$ and the between-class differences in $\hat{\theta}_{mi}$. Next, it will be examined what the specific effects are on the weighted and assigned latent trait estimates, and which of these two procedures performs best.

Further, it is examined whether the mixture IRT model improves the prediction compared to a one-class model. To approach population values, a data set of 25000 response patterns is generated for each condition. A one-class model is fitted to these data sets (Vermunt, 1997). The maximum likelihood estimates of the item parameters of the one-class model were situated approximately in between the parameter values of the two latent classes used in the simulation. Next, the item parameters of this one-class model are fixed for the original five hundred data sets of each condition to estimate the latent trait scores, assuming that this one-class model would fit the data.

5.2.2 Results

The mean Pearson correlation coefficients of the simulation are shown in Table 5.3, as well as the correlations between the class-specific latent trait estimates and the mean variance of the individual latent class probabilities. The differences in the class-specific latent trait estimates $\hat{\theta}_{mi}$ are expressed by the correlations between the latent trait estimates of the two latent classes. First, consider the short tests. For the Birnbaum model, the weighted

Table 5.3: *Mean Pearson Correlation Coefficients of the Weighted and Assigned Latent Trait Estimates with the Simulated Criterion and Latent Trait Values, as well as the Correlations Between the Class-specific Latent Trait Estimates $\hat{\theta}_m$ and the Mean Variance of the Individual Latent Class Probabilities*

Condition	Criterion		Trait			
	Weighted $\hat{\theta}_i$	Assigned $\hat{\theta}_i$	Weighted $\hat{\theta}_i$	Assigned $\hat{\theta}_i$	r $(\hat{\theta}_1\hat{\theta}_2)$	$\text{var}_m(\hat{\pi}_{mi})$
a1	0.360	0.360	0.800	0.800	1.0	0.0156
a2	0.355	0.352	0.787	0.782	0.940	0.0545
a3	0.346	0.341	0.770	0.757	0.721	0.1296
b1	0.360	0.360	0.800	0.800	1.0	0.0465
b2	0.354	0.351	0.788	0.782	0.940	0.0728
b3	0.347	0.341	0.771	0.758	0.721	0.1386
A1	0.397	0.397	0.882	0.882	1.0	0.0302
A2	0.394	0.392	0.875	0.872	0.960	0.1135
A3	0.391	0.389	0.869	0.864	0.774	0.2433
B1	0.397	0.397	0.882	0.882	1.0	0.0891
B2	0.393	0.392	0.875	0.873	0.960	0.1404
B3	0.392	0.390	0.870	0.865	0.774	0.2530

sum of the item responses, with weights equal to the item discrimination parameters, is a sufficient statistic for estimating the latent trait scores (Birnbaum, 1968). Therefore, the correlations between the class-specific latent trait estimates for Conditions a and b are the same. Differences in the difficulty parameters across latent classes have no influence on the class-specific latent trait estimates. In Condition 1, the discrimination parameters are equal across latent classes, and therefore produce identical latent trait estimates for both latent classes. The differences in $\hat{\theta}_{mi}$ and $\text{var}_m(\hat{\pi}_{mi})$ both increase as the differences in discrimination parameters across latent classes become larger. The differences in the difficulty parameters across latent classes have the same influence but only on $\text{var}_m(\hat{\pi}_{mi})$. This effect becomes smaller as the discrimination parameters vary increasingly.

Because for Condition 1, the latent trait estimates, $\hat{\theta}_{mi}$, are identical across latent classes, there are of course no differences between weighted and assigned latent trait estimates. As $\text{var}_m(\hat{\pi}_{mi})$ and the differences in $\hat{\theta}_{mi}$ increase, the correlations of the weighted and assigned latent trait estimates decrease, in favor of the weighting procedure. The weighted latent trait estimates show higher correlations with both the simulated criterion and latent trait values compared to the assigned latent trait estimates ($p < .01$). As

expected, the differences between weighting and assignment become larger when the class-specific latent trait estimates vary increasingly, as reflected by lower correlations between the latent trait estimates of the two latent classes. However, $\text{var}_m(\hat{\pi}_{mi})$ increased as well. This means that the effect of the class-specific latent trait estimates dominates the effect of $\text{var}_m(\hat{\pi}_{mi})$. It is shown that weighted latent trait estimates provide a more accurate prediction compared to assigned latent trait estimates.

Next, consider the results of Conditions A1 trough B3, for longer tests. The effects of differences in item parameters across latent classes are comparable to those of the shorter tests. As the discrimination parameters vary increasingly across latent classes, the correlations between the latent trait estimates of the two latent classes decrease, though the decrease is smaller than for the short test. The $\text{var}_m(\hat{\pi}_{mi})$ is clearly higher for the long than for the short tests. For longer tests, a specific response pattern becomes more typical for a particular latent class compared to a shorter response pattern, resulting in increasing $\text{var}_m(\hat{\pi}_{mi})$. Also, the number of subjects correctly allocated to the latent classes (as assumed in the simulation) increased as the differences between the IRT models and $\text{var}_m(\hat{\pi}_{mi})$ increased.

The correlations of the weighted and assigned latent trait estimates with both the simulated criterion and latent trait values are higher for the long tests compared to the short tests. This is expected as longer tests provide more reliable latent trait estimates. The weighting procedure performs better than assignment ($p < .01$). However, the differences in weighting and assignment are smaller than for the short tests, which suggests that the influence of $\text{var}_m(\hat{\pi}_{mi})$ starts to dominate the effect of differences in $\hat{\theta}_{mi}$ for longer tests. As the variances increase it was expected that the differences between weighting and assignment would decrease. Here, the variances are clearly higher, compared to the short tests. The larger variances show to reduce the magnitude of the differences between weighting and assignment.

The results show that both effects are indeed related to the performance of weighted and assigned latent trait estimates in predicting a simulated criterion and latent trait values. For small sets of items, the effect of the differences in class-specific latent trait estimates dominates the influence of the variances. The weighted latent trait estimates show higher correlations with the simulated criterion and latent trait values, compared to the assigned latent trait estimates. However, for the longer tests, the effect of the

variances starts to dominate the influence of the differences in $\hat{\theta}_{mi}$. The differences between weighting and assignment become smaller compared to the short tests, although still a weighted latent trait estimate performs best.

5.2.3 Prediction with a One-class model

Improvement in prediction with the mixture IRT model compared to the one-class model, should be indicated by lower correlations with the simulated criterion and latent trait values for this one-class model. The Pearson correlations of these latent trait estimates with the simulated criterion and latent trait values are given in Table 5.4. It is seen

Table 5.4: *Mean Pearson Correlations of the Latent Trait Estimate for the One-Class Model with the Simulated Criterion and Latent Trait Scores*

Condition	Criterion	Trait
a1	0.360	0.800
a2	0.354	0.785
a3	0.340	0.757
b1	0.360	0.799
b2	0.353	0.786
b3	0.341	0.758
A1	0.397	0.882
A2	0.392	0.873
A3	0.383	0.852
B1	0.397	0.882
B2	0.392	0.873
B3	0.383	0.852

that the conditions that only differ in the difficulty parameters across latent classes (for example, a1 and b1) have (approximately) equal correlations for the one-class model, which are consistent with those of the mixture IRT models. The correlations decrease as the discrimination parameters vary increasingly in the mixture IRT models. As the differences in the discrimination parameters increase, the weighted latent trait estimates show higher correlations than the latent trait estimates for the one-class model. All differences in correlations are significant ($p < .05$), except for the correlations with the simulated criterion in Condition b2. The assigned latent trait estimates only perform

better than the one-class model in Conditions A3 and B3 ($p < .01$), where the item parameters across the latent classes of the long tests vary most. For all other conditions the latent trait estimates of the one-class model perform equally well or better than the assigned latent trait estimates. The differences between the correlations are significant for Conditions a2 and b2 ($p < .01$), and Condition A2 (with respect to the simulated latent trait).

These results show that weighted latent trait estimates perform equally well or better than assigned latent trait estimates and latent trait estimates of the one-class model. The differences are influenced by both the differences in the class-specific latent trait estimates and the variance of the individual latent class probabilities. As the differences in $\hat{\theta}_{mi}$ increase the differences between weighting and assignment become larger. The differences between weighting and assignment decrease when $\text{var}_m(\hat{\pi}_{mi})$ increases. Both effects turn up when the discrimination parameters vary increasingly across latent classes. The number of items determines which of the two effects dominates the other. When compared to latent trait estimates of the one-class model, the weighted latent trait estimates provide the most accurate predictions of both the simulated criterion and latent trait values.

5.3 Empirical Example

The scales extraversion and neuroticism are analyzed, of the Amsterdam Biographical Questionnaire (Amsterdamse Biografische Vragenlijst, or ABV; Wilde, 1970), with 21 and 30 items respectively. The questionnaire is based on the MPI (Maudsley Personality Inventory; Eysenck, 1959), which is an early version of the EPI (Eysenck Personality Inventory; Eysenck & Eysenck, 1975). The items have three response categories; "yes", "?" and "no". The questionnaire was administered to 3943 applicants for blue collar jobs and lower administrative jobs at the Dutch Railways, regional bus companies and road transport companies in the Netherlands. In a previous study by Maij-de Meij, Kelderman and Van der Flier (in press), a mixture version of the nominal response model (mNRM) with three latent classes was found to be the best fitting model. This model will be used for further analyses. The latent classes differ with respect to ethnic background and social desirability, both incorporated as exogenous variables. The differences between the latent

classes were shown to be related to differential use of the response scale, such as the use of the "?" category. For a part of the sample, a criterion measure is available corresponding to the scales extraversion (N=344) and neuroticism (N=342). This is a judgement on a five-point scale about both extraversion and emotional stability by a psychologist based on a selection-interview. The psychologist had no knowledge of the test results of the applicant.

Individual latent class probabilities and MAP estimates $\hat{\theta}_{mi}$ are obtained, as described above, based on the parameters of the mixture NRM model found in a previous study (Maij-de Meij et al., in press). The latent trait estimates are transformed to be on the scale of the external criterion measure. To evaluate which of the four combinations of the procedures described earlier is the most accurate, the final latent trait estimates are related to the external criterion measure.

5.3.1 Results for Weighting and Assignment

Table 5.5 gives the mean polyserial correlations between the external criterion measure and both the weighted and assigned latent trait estimates across the two subsamples for the extraversion and neuroticism scale. The order of the correlations for the extraversion scale is as expected. The weighted latent trait estimates provide a more accurate prediction of the criterion compared to the assigned latent trait estimates. Also, the scale transformation Method II improves the prediction compared to Method I. For the neuroticism scale, the correlations are closer to each other. The assigned latent trait estimate, following scale transformation Method I, provides the most accurate prediction of the criterion. However, for both personality scales, the differences between the correlations are not significant.

The simulation study showed that both the differences in $\hat{\theta}_{mi}$ and $\text{var}_m(\hat{\pi}_{mi})$ may

Table 5.5: *Mean Polyserial Correlations for the Relation of the Criterion Measure with Four Latent Trait Estimates for the Extraversion and Neuroticism Scale*

	Scale transformation Method I		Scale transformation Method II	
	Weighted $\hat{\theta}_i$	Assigned $\hat{\theta}_i$	Weighted $\hat{\theta}_i$	Assigned $\hat{\theta}_i$
Extraversion	0.231	0.210	0.243	0.227
Neuroticism	0.344	0.349	0.336	0.341

influence the differences between weighting and assignment. The Pearson correlations between the class-specific latent trait estimates (after transformation using the regression parameters) for both personality scales were examined, see Table 5.6, to study whether this also affected the empirical results. The class-specific latent trait estimates of neuroticism

Table 5.6: *Mean Pearson Correlations for the Relations among the Transformed Class-specific Latent Trait Estimates for the Extraversion and Neuroticism Scale, as well as the Variance of the Individual Latent Class Probabilities*

	$r\ (\hat{\theta}_1\hat{\theta}_2)$	$r\ (\hat{\theta}_1\hat{\theta}_3)$	$r\ (\hat{\theta}_2\hat{\theta}_3)$	$\text{var}_m(\hat{\pi}_{mi})$
Extraversion	0.823	0.650	0.630	0.174
Neuroticism	0.891	0.850	0.843	0.191

are more similar compared to the extraversion scale. This can be concluded from the higher correlations among the class-specific latent trait estimates. When the differences in $\hat{\theta}_{mi}$ increase, as is seen for the extraversion scale, the differences between weighting and assignment increase as well. This is in agreement with the results of the simulation study.

Furthermore, the variances of the individual latent class probabilities are examined, see Table 5.6. For the extraversion scale, $\text{var}_m(\hat{\pi}_{mi})$ is smaller than for the neuroticism scale. The simulation study showed that increasing variances were associated with a decrease in the differences between weighting and assignment (when comparing short and large tests). The neuroticism scale, with 30 items, shows a higher variance compared to the extraversion scale, with 21 items. This again may explain the larger differences between weighting and assignment for the extraversion scale, compared to the neuroticism scale.

The simulation study showed that differences in $\hat{\theta}_{mi}$ and $\text{var}_m(\hat{\pi}_{mi})$ were related to differences in the item parameters, specifically the discrimination parameters. Inspecting the differences in the discrimination parameters across latent classes for both personality scales shows that for the extraversion scale the differences between the parameters are larger compared to the neuroticism scale.

5.3.2 Prediction with a One-class Model

The performance of a one-class model was examined for the empirical example as well. A one-class model was fitted to the data, where the estimated item parameters were

subsequently used to estimate latent trait values. The correlations between the latent trait estimates for the one-class model with the criterion measures were .284 and .283 for the extraversion and neuroticism scale respectively. This means that for the extraversion scale the one-class model performs better than the weighted and assigned latent trait estimates. However, the differences between these correlations are not significant. For the neuroticism scale, the one-class model performs worse compared to the mixture IRT model. These differences are significant for the comparisons of the trait estimates of the one-class model and the weighted latent trait estimates for both equating methods, z = 2.03,$p = .04$ and $z = 2.02, p = .04$ for scale transformation Method I and II respectively.

5.4 Conclusion

This study demonstrates that the assignment procedure may not always provide the best latent trait estimates. The proposed weighting procedure performed in most cases equally well or better than the assignment procedure. These results are found under the assumption that the same construct is measured in each of the latent classes. On the whole, the differences between the correlations are quite small.

In the simulation study, the specific effects of differences in item parameters across latent classes on the prediction of a simulated criterion and latent trait values were examined. It was shown that the differences between weighting and assignment increase as the differences in the discrimination parameters across latent classes become larger. The weighted latent trait estimates showed higher correlations with both the simulated criterion and latent trait values compared to the assigned latent trait estimates. The effects of the item parameters could be explained by their influence on the differences in $\hat{\theta}_{mi}$ and $\text{var}_m(\hat{\pi}_{mi})$. As the discrimination parameters varied increasingly, the differences in $\hat{\theta}_{mi}$ as well as $\text{var}_m(\hat{\pi}_{mi})$ increased. For the short test, the differences between weighting and assignment increased, in favor of the weighting procedure, which implies that the effect of differences in $\hat{\theta}_{mi}$ dominates the effect of the increasing $\text{var}_m(\hat{\pi}_{mi})$. The differences between weighting and assignment were smaller for the longer tests. There the influence of the increasing variances started to dominate the effect of differences in $\hat{\theta}_{mi}$.

It can be concluded that computing a weighted latent trait estimate should be

preferred to an assigned latent trait estimate. The weighting procedure performs equally well or better than the assigned latent trait estimates, in predicting the simulated criterion and latent trait values. Correlations for a one-class model showed that weighted latent trait estimates should be preferred over a one class model as well. The assigned latent trait estimates did not always perform better than the estimates of the one-class model. In particular for shorter tests, the weighting procedure performs better.

In the empirical example of two personality scales, latent trait estimates were transformed to be on the same scale across latent classes. Under the assumption that the same trait is measured in each latent class, not only within but also between class comparison is possible. It was shown that for the extraversion scale a weighted latent trait estimate provided a better prediction of an external criterion, than an assigned latent trait estimate, though both performed worse than the one-class model. These differences did not prove to be significant. For the neuroticism scale differences in predictive validity were small. The assigned latent trait estimate, following scale transformation Method I, gave the best prediction. Both methods clearly improved the prediction compared to the one-class model, although only the correlations for the weighted latent trait estimates were significantly higher.

The results were in agreement with the explanations found in the simulation study. For the extraversion scale $\text{var}_m(\hat{\pi}_{mi})$ was smaller, and the differences in $\hat{\theta}_{mi}$ were larger, which both may explain the larger differences between weighting and assignment. Also, the influence of the length of the scales was in agreement with the simulation study. The neuroticism had almost 50 percent more items, which was shown to be related to smaller differences between weighting and assignment.

It has been shown that a weighted latent trait estimate should be considered as an alternative to assignment, when applying mixture IRT models. In particular for short tests, it is preferred to compute weighted latent trait estimates and improve the estimation of the latent trait and prediction of external criteria.

Chapter 6

Conclusion and Discussion

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

6.1 Summary

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

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

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

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

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

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

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

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

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

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

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

estimates of latent trait values and prediction of external criteria.

6.2 Further Research

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

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

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

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

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

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

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

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

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

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

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

Bibliography

Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships.* New York: Holt, Rinehart and Winston.

Anderson, C. J., & Vermunt, J. K. (2000). Log-multiplicative associations models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, *30*, 81-121.

Angoff, W. H. (1993). Perspectives on differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*, 1235-1245.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*, 491-509.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, *81*, 261-272.

Bass, B. M. (1957). Faking by sales applicants of a forced choice personality inventory. *Journal of Applied Psychology*, *41*, 403-404.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 506-520.

Berdie, R. F. (1961). Intra-individual variability and predictability. *Educational and Psychological Measurement*, *3*, 663-676.

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement*, *13*, 164-169.

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

Blom, H. (1992). *Feedback exposure questionnaire.* Haarlem: Blom and Partners.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381-409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis.* New York: Wiley.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks: Sage.

Carver, C. S., & Scheier, M. F. (1995). *Perspectives on personality.* Boston: Allyn and Bacon.

Clogg, C. C. (1982). Using association models in sociological research: Some examples. *American Journal of Sociology*, *88*, 114-134.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133-148.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research and Practice*, *20*, 225-233.

Colvin, C. R., & Longueuil, D. (2001). Eliciting self-disclosure: The personality and behavioral correlates of the opener scale. *Journal of Research in Personality*, *35*, 238-246.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation.* London: Chapman and Hall.

Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, *30*, 171-178.

Cozby, P. C. (1973). Self-disclosure: A literature review. *Psychological Bulletin*, *79*, 73-91.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243-276.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories from dimensions. *Psychological Review*, *112*, 129-158.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

DuBois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, *35*, 869-884.

Dunnett, S., Koun, S., & Barber, P. J. (1981). Social desirability in the Eysenck Personality Inventory. *British Journal of Psychology*, *72*, 19-26.

Dunnette, M. D., MacCartney, K., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology*, *15*, 13-24.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York, N.Y.: Dryden press.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20-30.

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (p. 255-270). New York: Springer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Eysenck, H. J. (1959). *Manual of the Maudsley Personality Inventory.* London: University of London Press.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* London: Hodder & Stoughton.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review*, *92*, 317-349.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, *40*, 374-377.

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, *47*, 81-86.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement*, *28*, 173-189.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, *70*, 755-768.

Goodman, L. A. (1979). Simple models for the anaysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, *74*, 537-552.

Goodman, L. A. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*, 1-22.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 415-441.

Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stoufler, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II, Vol. IV: Measurment and prediction* (p. 60-90). Princeton: Princeton University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, *69*, 192-203.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: Sage.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology*, *89*, 687-699.

Hoekstra, H. A., Ormel, J., & De Fruyt, F. (1995). *NEO-Personality Inventory.* Lisse: Swets Test Publishers.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale. *Educational and Psychological Measurement*, *67*, 280-299.

Hui, C. H., & Triandis, H. C. (1989). Effect of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*, 296-309.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*, 243-252.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*, 321-357.

Johnson, T., Kulesa, T., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264-277.

Jourard, S. M. (1971). *The transparant self.* New York: Van Nostrand.

Jourard, S. M., & Lasakow, P. (1958). Some factors in self-disclosure. *Journal of Abnormal and Social Psychology*, *56*, 91-98.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*, 681-697.

Kelderman, H. (2007). Loglinear multivariate and mixture Rasch models. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (p. 77-97). New York: Springer.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*, 307-327.

Kelderman, H., & Molenaar, P. C. M. (2007). The effect of individual differences in factor loadings on the standard factor model. *Multivariate Behavioral Research*, *42*, 435-456.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*, 261-276.

Koeller, O. (1994, april). *Identification of guessing behavior on the basis of the mixed Rasch model.* (Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA)

Kok, F. G., Mellenbergh, G. J., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295-303.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Kolk, N. J., Born, M. P., & Van der Flier, H. (2004). Three method factors explaining the low correlations between assessment center dimension ratings and scores on personality inventories. *European Journal of Personality*, *18*, 127-141.

Langeheine, R., & Rost, J. (1988). *Latent trait and latent class models.* New York: Plenum Press.

Lanning, K. (1988). Individual differences in scalability: An alternative conception of consistency for personality theory and measurement. *Journal of Personality and Social Psychology*, *55*, 142-148.

Lauritzen, S. L. (1996). *Graphical models.* Oxford: Clarendon Press.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, A. E. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

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

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

Maij-de Meij, A. M., Kelderman, H., & Van der Flier, H. (2005). Latent-trait latent-class analysis of self-disclosure in the work environment. *Multivariate Behavioral Research*, *40*, 435-460.

Maij-de Meij, A. M., Kelderman, H., & Van der Flier, H. (in press). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences in degree and differences in kind. *Journal of Personality*, *60*, 117-174.

Meiser, T., Hein-Eggers, M., Rompe, P., & Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models : A comparative and integrative approach. *Applied Psychological Measurement*, *19*, 377-391.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

Miller, L. C., Berg, J. H., & Archer, R. L. (1983). Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology*, *44*, 1234-1244.

Miller, L. C., & Read, S. J. (1987). Why am I telling you this? Self-disclosure in a goal-based model of personality. In V. J. Derlega & J. H. Berg (Eds.), *Self-disclosure: Theory, research and therapy* (pp. 35–58). New York: Plenum.

Mischel, W. (1968). *Personality and assessment.* Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Morris, L. W. (1979). *Extraversion and introversion: An interactional perspective.* Washington: Hemisphere Publishing Corporation.

Morton, T. L. (1978). Intimacy and reciprocity of exchange: A comparison of spouses and strangers. *Journal of Personality and Social Psychology*, *36*, 72-81.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Omarzu, J. (2000). A disclosure decision model: Determining how and when individuals will self-disclose. *Personality and Social Psychological Review*, *4*, 174-185.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598-609.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleiter & J. A. Wiggings (Eds.), *Personality assessment via questionnaires* (p. 143-165). Berlin: Springer-Verlag.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.

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

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151.

Rijkes, C. P. M., & Kelderman, H. (2007). Latent response Rasch models for strategy shifts in problem solving. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (p. 316-333). New York: Springer.

Rost, J. (1990). Rasch models in latent classes: An intergration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

Rost, J., Carstensen, C., & Von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (p. 324-332). Münster [etc.]: Waxmann.

Rubin, Z. (1975). Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology*, *11*, 233-260.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 281-315). Hillsdale, NJ: Lawrence Erlbaum.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Slobin, D. I., Miller, S. H., & Porter, L. W. (1968). Forms of address and social relations in a business organization. *Journal of Personality and Social Psychology*, *8*, 289-293.

Smit, A., Kelderman, H., & Van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online*, *4*, 19-32.

Smit, A., Kelderman, H., & Van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, *5*, 31-43.

Smit, A., Kelderman, H., & Van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, *8*, 23-50.

Steel, J. L. (1991). Interpersonal correlates of trust and self-disclosure. *Psychological Reports*, *68*, 1319-1320.

Te Nijenhuis, J., & Van der Flier, H. (1999). Bias research in The Netherlands: Review and implications. *European Journal of Psychological Assessment*, *15*, 162-175.

Te Nijenhuis, J., Van der Flier, H., & Van Leeuwen, L. (1997). Comparability of personality test scores for immigrants and majority group members: Some Dutch findings. *Personality and Individual Differences*, *23*, 849-859.

Thissen, D., Chen, W. H., & Bock, R. D. (2003). *Multilog 7.0 [Computer program].* Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67-114). Hillsdale, NJ: Lawrence Erlbaum.

Van de Vijver, F. J. R., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review*, *52*, 215-236.

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

Van Hemert, D. A., Baerveldt, C., & Vermande, M. (2001). Assessing cross-cultural item bias in questionnaires: Acculturation and the measurement of social support and family cohesion for adolescents. *Journal of Cross-Cultural Psycholgy*, *32*, 381-396.

Vansteelandt, K., & Van Mechelen, I. (2004). The personality triad in balance: Multidimensional individual differences in situation-behavior profiles. *Journal of Research in Personality*, *38*, 367-393.

Vermunt, J. K. (1997). *ℓEM: A general program for the analysis of categorial data [Computer program].* Tilburg, The Netherlands: Tilburg University. (Internet: www.uvt.nl/faculteiten/fsw/organisatie/departementen/ mto/software2.html)

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

Von Davier, M., & Rost, J. (1997). Self-monitoring - a class variable? In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (p. 296-304). Münster [etc.]: Waxmann.

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the Generalized Partial-Credit model. *Applied Psychological Measurement*, *28*, 389-406.

Vonk, R. (1999). *Cognitieve sociale psychologie: Psychologie van het dagelijkse denken en doen.* Utrecht: Uitgeverij Lemma BV.

Wilde, G. (1970). *Neurotische labiliteit gemeten volgens de vragenlijstmethode [neurotic lability measured by the questionnaire method].* Amsterdam: Van Rossen.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*, 276-289.

Yamamoto, K. (1989). *A HYBRID model of IRT and latent class models* (Tech. Rep. No. ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995). *Estimating the effect of test lenght and test time on parameter estimation using the HYBRID model* (Tech. Rep. Nos. TOEFL Research Report 10, ETS Research Report RR-95-2). Princeton, NJ: Educational Testing Service.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model Item Response Theory. *Organizational Research Methods*, *7*, 168-190.

Samenvatting

In de sociale en gedragswetenschappen zijn veel attributen niet direct observeerbaar. Op basis van geobserveerde responses op bijvoorbeeld tests en vragenlijsten worden uitspraken gedaan over cognitieve capaciteiten, persoonlijkheid en attitudes. Met behulp van modellen met latente variabelen worden de associaties tussen de manifeste, geobserveerde, variabelen verklaard. Er is een algemeen raamwerk van modellen ontwikkeld waarbij de kans op een item response werd gerelateerd aan een onderliggende latente variabele. Deze latente variabele kan continu zijn of discreet. In de item response theorie (IRT) beschrijft een continue latente-trek variabele de kwantitatieve verschillen tussen personen. Kwalitatieve verschillen worden in kaart gebracht door een discrete latente-klasse variabele. Het centrale thema in dit proefschrift is de toepassing van mixture IRT modellen. Deze modellen voegen de latente-trek en latente-klasse variabele in één model samen. Er wordt een aantal homogene subgroepen, latente klassen, onderscheiden. Binnen deze latente klassen zijn verschillende latente-trek modellen van toepassing, in die zin dat de item parameterschattingen over klassen mogen verschillen. De mixture IRT modellen beschrijven de heterogeniteit binnen en tussen groepen. Op deze manier is het mogelijk om gelijktijdig te kijken naar kwantitatieve en kwalitatieve verschillen in antwoordgedrag van personen.

In Hoofdstuk 1 werd een aantal interpretaties van kwalitatieve verschillen gegeven. Er kunnen kwalitatieve verschillen voorkomen in het gemeten construct, zoals de situationele bepaaldheid van een eigenschap. Kwalitatieve verschillen kunnen ook ongerelateerd aan het gemeten contruct zijn. Echter, invloeden op het antwoordgedrag die niet direct gerelateerd zijn aan de latente trek kunnen mogelijk wel inhoudelijk interessant zijn. Denk hierbij bijvoorbeeld aan antwoordtendensen en sociale wenselijkheid.

In Hoofdstuk 2 werd het construct "self-disclosure" nader onderzocht. Self-disclosure heeft betrekking om het delen van informatie over jezelf met anderen. Een mixture IRT

model met drie latente klassen werd geïdentificeerd als best passende model. De latente trek gaf de algemene neiging weer om meer of minder self-disclosure te tonen. De personen in de latente klassen verschilden in hun patroon van self-disclosure, afhankelijk van de ontvanger van de disclosure. Dit betekent dat self-disclosure situatie afhankelijk is, waarbij de situatie betrekking heeft op de persoon naar wie de disclosure plaats vindt. Hiermee werd het concept selectiviteit in self-disclosure geïntroduceerd. Het onderzoek heeft laten zien hoe mixture IRT modellen inzicht kunnen geven in de situationele bepaaldheid van persoonlijkheidseigenschappen.

Vervolgens werd in Hoofdstuk 3 een latente DIF-detectie methode vergeleken met een manifeste DIF-detectie methode. Er werd gekeken hoe de twee methoden presteerden wat betreft het identificeren van DIF items. De latente DIF-detectie methode gaat uit van het fitten van een mixture IRT model, waarna wordt gekeken welke item parameters significant verschillen over klassen. De gangbare manifeste DIF-detectie methoden onderzoeken het verschillend functioneren van items over manifeste groepen. Het voordeel van de latente DIF-detectie methode is dat vooraf geen aannames gedaan hoeven te worden over de bron van de bias. Echter, een manifeste variabele kan wel worden opgenomen in het model als indicator voor de latente klasse variabele. Dit biedt tevens de mogelijkheid om de relatie met de latente-klasse variabele te onderzoeken.

Met een simulatie studie werd de identificatie van DIF items onderzocht met beide DIF-detectie methoden. Gevarieerd werden steekproefgrootte en de sterkte van de relatie tussen de bron van de bias en de manifeste variabele. De resultaten lieten zien dat de manifeste DIF-detectie methode minder goed werkte naarmate de relatie zwakker werd tussen de manifeste variabele en de bron van de bias. De latente DIF-detectie methode deed het goed, ook in de condities waarin de relatie tussen de manifeste variabele en de bron van de bias zwakker was. De latente DIF-detectie methode, zonder manifeste variabele als indicator voor de latente-klasse variabele, deed het alleen beter in de conditie waar de correlatie tussen de bron van de bias en de manifeste variabele nul was. In de andere condities deed de latente DIF-detectie methode met de manifeste indicator het beter. De studie laat zien hoe mixture IRT modellen perspectieven bieden voor het verbeteren van DIF-detectie, met als bijkomend voordeel dat vooraf geen aannames gedaan hoeven te worden over de bron van de bias.

Kwalitatieve verschillen tussen de latente klassen kunnen ook optreden door een

verschillende betekenis van de response categorieën van de items. In Hoofdstuk 4 werden de schalen extraversie en neuroticisme onderzocht afkomstig van de Amsterdamse Biografische Vragenlijst (Wilde, 1970). Aansluitend bij onderzoek van Smit, Kelderman en Van der Flier (2003) werd er in deze studie een grotere dataset onderzocht, en werden geen restricties gelegd op het aantal te identificeren latente klassen. Voor beide persoonlijkheidsschalen werden drie latente klassen gevonden. In overeenstemming met eerder onderzoek hadden de verschillen tussen de klassen allereerst te maken met het verschillend gebruik van de "?" categorie. Personen uit een van de latente klassen leken de response categorie te vermijden. Deze latente klasse kon tevens gekarakteriseerd worden door een lage neiging tot sociaal wenselijk antwoorden. Met name voor de neuroticisme schaal was een duidelijke voorkeur voor de "?" categorie waar te nemen voor personen uit een van de latente klassen. Deze latente klasse werd verder gekenmerkt door personen uit ethnische minderheidsgroepen, en een hogere neiging tot sociaal wenselijk antwoorden. Tenslotte bestond de derde latente klasse met name uit autochtone nederlanders met een hogere neiging sociaal wenselijk te antwoorden.

Een belangrijke vraag was of mixture IRT modellen ook gebruikt kunnen worden om een betere voorspelling mogelijk te maken van iemands criterium score. Voor een deel van de steekproef in Hoofdstuk 4 waren criterium data aanwezig, die overeen kwamen met respectievelijk extraversie en emotionele stabiliteit. De voorspellende waarde van het mixture IRT model werd vergeleken met de voorspelling van het criterium onder de aanname van een een-klasse model. Uit de resultaten konden geen eenduidige conclusies getrokken worden wat betreft de waarde van mixture IRT modellen voor predictie van een extern criterium. Voor de extraversie schaal liet de trekschatting onder het een-klasse model hogere correlaties met het criterium zien dan de trekschatting onder het mixture IRT model. Voor de neuroticisme schaal lagen de resultaten andersom. Trekschattingen onder het mixture IRT model voorspelden het criterium beter dan schattingen onder aanname van een een-klasse model.

In Hoofdstuk 5 werd dieper ingegaan op de vraag of mixture IRT modellen een betere voorspelling van een criterium score mogelijk maken. Over het algemeen worden personen toegewezen aan latente klassen op basis van de hoogste klasse kans gegeven het antwoordpatroon van de persoon. Vervolgens wordt de latente trekwaarde toegewezen, die voor die persoon voor de betreffende latente klasse geschat is. De zekerheid van toewijzing

aan een latente klasse kan echter sterk verschillen over personen. Daarom werd als alternatief een gewogen trekschatting voorgesteld. De klasse-specifieke trekschattingen werden in dat geval gewogen met de corresponderende latente-klasse kansen. In geval van hoge zekerheid van toewijzing (hoge klasse kans) zou het verschil tussen wegen en toewijzen klein moeten zijn. Dit verschil zou moeten toenemen naarmate de klasse kansen meer gelijk over de klassen verdeeld zijn, en/of wanneer de klasse-specifieke trekschattingen meer van elkaar verschillen over klassen.

Uit een simulatie studie bleek dat een gewogen latente trekschatting een hogere correlatie met de gesimuleerde latente trek en criterium score had dan de toegewezen latente trekschatting. In vergelijking met de gewogen latente trekschattingen, leverden trekschattingen op basis van een een-klasse model over het algemeen lagere correlaties op met gesimuleerde latente trek en criterium scores. Op basis van de resultaten onder verschillende condities werd geconcludeerd dat een gewogen trekschatting vooral beter voorspelt bij kortere tests en wanneer er grote verschillen in item parameters tussen de klassen zijn. Voor de twee persoonlijkheidsschalen uit Hoofdstuk 4 werden eveneens gewogen en toegewezen trekwaarden geschat. Hiervoor werden de klasse-specifieke trekschattingen eerst op eenzelfde schaal over klassen gebracht. Voor de extraversie schaal waren de correlaties van het criterium met de gewogen latente trekschattingen hoger dan de correlaties van het criterium met de toegewezen trekschattingen. Beide waren echter lager dan de trekschattingen op basis van een een-klasse model. Voor de neuroticisme schaal lagen de correlaties van de gewogen en toegewezen latente trekschattingen dicht bij elkaar. Alle correlaties op basis van het mixture IRT model waren hoger dan de trekschattingen op basis van het een-klasse model.

Kwalitatieve verschillen zijn vanuit verschillende perspectieven besproken, met name gericht op situationele bepaaldheid, DIF en antwoordtendensen. Verschillen in manifestaties van persoonlijkheidseigenschappen zijn in kaart te brengen. Dit is getoond voor self-disclosure. Het toepassen van mixture IRT modellen geeft tevens mogelijkheden om invloeden in kaart te brengen die niet direct betrekking hebben op de te meten latente trek. De resultaten van de latente DIF-detectie methode bieden goede perspectieven voor toekomstig onderzoek. Conclusies over de waarde van mixture IRT modellen als methode om schattingen van de latente trek en voorspelling van criterium scores te verbeteren kunnen nog niet eenduidig getrokken worden.