Applications of Item Response Theory to Non-Cognitive Data

Iris J.L. Egberink
The research presented in this dissertation was supported by funding from PiCompany BV. The opinions expressed by authors are their own and do not necessarily reflect the views of PiCompany BV.

© 2010 Applications of Item Response Theory to Non-Cognitive Data, Iris J. L. Egberink, University of Groningen

Cover designed by Huub A. F. Mulhof
Layout by Iris J. L. Egberink
Printed by Ipskamp Drukkers B.V., Enschede
Applications of Item Response Theory to Non-Cognitive Data

Proefschrift

ter verkrijging van het doctoraat in de Gedrags- en Maatschappijwetenschappen aan de Rijksuniversiteit Groningen

op gezag van de Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op donderdag 25 november 2010 om 13.15 uur

door

Iris Jo-Anne Lieneke

geboren op 14 september 1984 te Hengelo
Promotor: Prof. dr. R.R. Meijer

Copromotor: Dr. ir. B.P. Veldkamp

Beoordelingscommissie: Prof. dr. M.Ph. Born
Prof. dr. H.A. Hoekstra
Prof. dr. K. Sijtsma
# Contents

1 Introduction ...................................................................................................1  
1.1 Item Response Theory .................................................................................1  
1.2 Outline of this Thesis .................................................................................3  

2 An Item Response Theory Analysis of Harter’s Self-Perception Profile for  
Children or Why Strong Clinical Scales Should Be Distrusted .....................5  
2.1 Introduction ...............................................................................................6  
2.2 Method .......................................................................................................8  
2.3 Results ......................................................................................................12  
2.4 Discussion ...............................................................................................20  
2.5 Recommendations ...................................................................................22  

3 Detection and Validation of Unscalable Item Score Patterns Using Item  
Response Theory: An Illustration with Harter’s Self Perception Profile for  
Children ...........................................................................................................25  
3.1 Introduction ..............................................................................................26  
3.2 Item Response Theory and Individual Score Patterns ..........................27  
3.3 Method .....................................................................................................33  
3.4 Results ......................................................................................................36  
3.5 Discussion ................................................................................................46  
3.6 Some Practical Considerations for the Clinician ......................................48  
3.7 Appendix: Observations During Test Administration .............................49  

4 Detection of Aberrant Item Score Patterns in Computerized Adaptive  
Testing: An Empirical Example Using the CUSUM .................................. 51  
4.1 Introduction ..............................................................................................52  
4.2 Method .....................................................................................................53  
4.3 Results ......................................................................................................58  
4.4 Discussion ................................................................................................62
Chapter 1

Introduction

Tests and questionnaires play a crucial role in psychological assessment. Both cognitive measures (e.g., intelligence tests) and non-cognitive measures (e.g., mood questionnaires, personality questionnaires) belong to the practitioner’s toolkit in different fields of psychology. For example, in personnel selection procedures besides intelligence testing, often personality questionnaires are used to assess whether a candidate is suited for a particular job. Also, in the clinical field both cognitive and non-cognitive measures are used for diagnostic purposes and to select the most appropriate treatment for the diagnosed disorder. Because psychological tests and questionnaires are used to make important decisions, high-quality standards for the construction and the evaluation of these instruments are necessary. One of these standards is item response theory (IRT, e.g., Embretson & Reise, 2000; van der Linden & Hambleton, 1997). Although there has been no shortage of researchers demonstrating the potential of IRT in the cognitive domains, its use in the non-cognitive measurement area (e.g., personality, attitude, and psychopathology) has lagged behind that of other areas. Nevertheless, there are signs that applied researchers are beginning to use IRT with greater frequency in recent years (Reise & Waller, 2009; Weekers, 2009). The overarching aim of this thesis is to further contribute to the use of IRT in the non-cognitive domain.

1.1 Item Response Theory

IRT models are based on the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of items (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002). IRT explains the structure in the manifest responses by assuming the existence of a latent trait, denoted by the Greek letter \( \theta \). By means of IRT models it is possible to locate a person’s \( \theta \) and the characteristics of the items that make up the measurement instrument, on the same metric (i.e.,
latent trait continuum). The goal of fitting an IRT model is to identify an item response function (IRF) that describes the relation between \( \theta \) and the probability of responding to a specified response category. IRT models can be applied to both dichotomous (e.g., true-false) and polytomous data (e.g., 5-point Likert scale).

Most IRT models assume unidimensionality and a specified form for the IRF that can be checked empirically. Unidimensionality means that all items in the test measure the same latent trait with the result that persons can be ordered on a linear scale. Related to unidimensionality is the assumption of local independence, which holds that the responses in a test are statistically independent conditional on \( \theta \). Furthermore, it is assumed that the probability of endorsing an item is monotonically nondecreasing in \( \theta \).

There are parametric and nonparametric IRT models. Both types of models assume unidimensionality and local independence, but differ in the assumption regarding the form of the IRF. For parametric IRT models the IRF is defined by a parametric function, whereas for nonparametric IRT models the IRF is not defined by a parametric function, the only restriction regarding the form of the IRF is that it is nondecreasing. Recently, several authors have introduced and discussed the advantages of applying IRT models to construct personality scales and to explore the structure of personality data sets. For example, Waller, Tellegen, McDonald, and Lykken (1996) contrasted the use of IRT with principal component factor analysis, and Reise and Waller (2003) discussed the choice of an IRT model to analyze psychopathology test data. That is, they compared the fit of different parametric IRT models on 15 unidimensional factor scales from the Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A; Butcher et al., 1992). Most studies apply parametric IRT models to investigate the quality of personality and psychopathology tests (e.g., Panter, Swygert, Dahlstrom, & Tanake, 1997; Robie, Zickar, & Schmit, 2001; Steinberg, 1994; Waller, Thompson, & Wenk, 2000). However, Meijer and Baneke (2004), and Wismeijer, Sijtsma, van Assen, and Vingerhoets (2008) provided interesting applications of nonparametric IRT models to non-cognitive data.

Besides a more thorough analysis at the item level to evaluate the quality of a measurement instrument, IRT models also provide useful tools to evaluate the measurement accuracy and validity of item scores at the individual level. Based on an IRT model, it is possible to predict a person’s response behavior when confronted with a particular set of items and unlikely item score patterns can be identified (using so-called person-fit statistics), or groups of persons can be identified with similar answering behavior (e.g., Emons, 2008; Glas & Dagooy, 2007; Meijer & Sijtsma,
Although recently many methods have been proposed to identify persons with particular response behavior (for an overview see Meijer & Sijtsma, 2001), the practical usefulness of these methods has sometimes be questioned (e.g., Rudner, Bracey, & Skaggs, 1996).

1.2 Outline of this Thesis

This thesis demonstrates how IRT-based methods can be of help to improve measurement at the individual level. Chapter 2 illustrates how nonparametric and parametric IRT models can be used to evaluate the psychometric properties of questionnaires and inventories. In this chapter IRT models are applied to an often used inventory to measure self-concept: Harter’s (1985) Self-Perception Profile for Children (SPPC; Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004). Specifically, it is assessed which items in each scale mainly determine the construct that is being measured, and whether the scales can reliably distinguish persons across different values of the latent trait scale. The usefulness of some standard nonparametric tools is also evaluated.

Chapter 3 and 4 illustrate the usefulness of the person-fit methodology for non-cognitive measures. By means of an IRT model it is possible to predict a person’s response behavior when confronted with a particular set of questionnaire items. In person-fit research observed and expected item scores are compared and unlikely, aberrant, or inconsistent patterns are identified. However, in clinical practice and applied research, the fundamental question often is not whether unexpected item score patterns exist but whether the patterns have any theoretical or applied validity. In Chapter 3, the usefulness of person-fit methodology to identify aberrant response patterns (i.e., invalid test scores) using the same data as in Chapter 2 is explored. Information from person-fit statistics is combined with auxiliary information from personality theory and a person’s personal history. By means of these additional data more insight into possible causes that underlie inconsistent response behavior is obtained. In Chapter 4 the application of a relatively new method, the cumulative sum procedure (CUSUM; van Krimpen-Stoop & Meijer, 2002) to detect inconsistent response patterns in a computerized adaptive test for personality is illustrated. Combined information from the CUSUM, other personality measures, and interviews is used to interpret the test scores.

In personality assessment often self-report inventories are used to map personality characteristics. The use of self-report inventories is, however, not without problems. An often encountered problem is that people might interpret
subsets of items or response categories differently, which may affect the individual classification and the predictive validity of individual test scores (e.g., Austin, Deary, & Egan, 2006). In Chapter 5 a mixture IRT model (e.g., Rost, 1990; Rost & Langeheine, 1997) is applied to a Conscientiousness scale in a career development context to assess differences in response scale usage at the group level and to compare the predictive validity of the mixture IRT trait estimates with the unidimensional IRT trait estimates. Furthermore, differences in response behavior are explained using psychological theory.

An often encountered problem that may invalidate test scores on self-report personality inventories is faking or intentional response distortion. Research has shown that persons are able to significantly distort their answers on a wide variety of personality measures. In Chapter 6 the usefulness of different types of validity indicators to detect different types of aberrant response behavior is investigated in an external application group, an internal application group, and a career development group. Finally, the practical implications of the use of different validity indicators in psychological assessment are discussed.

The chapters in this thesis are self-contained, hence they can be read separately. Therefore, some overlap could not be avoided.
Chapter 2

An Item Response Theory Analysis of Harter’s Self-Perception Profile for Children or Why Strong Clinical Scales Should Be Distrusted

Abstract
We investigated the psychometric properties of the subscales of the Self-Perception Profile for Children with item response theory (IRT) models using a sample of 611 children. Results from a nonparametric Mokken analysis and a parametric IRT approach for boys (n = 268) and girls (n = 343) were compared. We found that most scales formed weak scales and that measurement precision was relatively low and only present for latent trait values indicating low self-perception. The subscales Physical Appearance and Global Self-Worth formed one strong scale. Children seem to interpret Global Self-Worth items as if they measure Physical Appearance. Furthermore, we found that strong Mokken scales (such as Global Self-Worth) consisted mostly of items that repeat the same item content. We conclude that researchers should be very careful in interpreting the total scores on the different Self-Perception Profile for Children scales. Finally, implications for further research are discussed.

This chapter has been accepted for publication as:
2.1 Introduction

A homogeneous, Guttman-type test yields high precision, but covers little ground. (...) It is worth sacrificing fidelity to attain (...) bandwidth (Cronbach, 1954, pp. 268-270)

Self-perception or self-concept is an important construct in developmental psychology, although there is still a debate about the exact relation between self-perception and important outcome variables. Self-perception is often used to refer to the way people evaluate their various abilities and attributes (e.g., Harter, 1985). Some authors claim that positive feelings about the self are important for healthy developmental outcomes, such as subjective well-being (e.g., DeNeve & Cooper, 1998; Diener & Diener, 1995), healthy social relationships and attachment (e.g., Leary, Tambor, Terdal, & Downs, 1995; Murray, Holmes, & Griffin, 2000), and academic achievement and occupational success (e.g., Elliott, 1996; Hansford & Hattie, 1982; Judge & Bono, 2001). Negative feelings of the self are claimed to be related to problematic outcomes, such as poorer mental and physical health (e.g., Roberts, Gotlib, & Kassel, 1996; Trzesniewski et al., 2006) and antisocial behavior (e.g., Donnellan, Trzesniewski, Robins, Moffitt, & Caspi, 2005; Rosenberg, Schooler, & Schoenbach, 1989). Recent evidence for the importance of self-concept is presented in, for example, Montague, Enders, Dietz, Dixon, and Morrison Cavendish (2008), who found a strong relation between depressive symptoms and self-perception in a school-based sample at risk for developing emotional and behavioral disorders. They found that when depressive symptoms decreased self-perception improved. Also, a low self-concept may be related to the development of obesity (e.g., Schumann et al., 1999).

Baumeister, Campbell, Krueger, and Vohs (2003) reviewed the self-esteem literature and concluded that self-esteem decreases the chances to become depressed and to develop eating disorders and is positively associated with life satisfaction. However, they found that it is unclear whether self-esteem determines positive social relations, academic achievement, and academic success.

Dubois and Tevendale (1999; see also Kernis, 2006) presented a critical review of the self-concept and self-esteem debate, where two opposite views were discussed, namely self-esteem as a “vaccine”, that is, as a construct that prevents against educational failure and crime, or as a “epiphenomenon”, that is, as a construct that indicates that “the interaction with the world is not going well”. Dubois and Tevendale (1999) concluded that self-concept could be investigated best when considered consisting of different distinct concepts and that the moderating
An Item Response Theory Analysis of Harter’s SPPC

influence of youth characteristics and environmental processes should be taken into account.

Harter’s (1985) Self-Perception Profile for Children (SPPC; Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004) is a popular instrument to measure self-concept. It is often used as a measure to determine self-concept in schools and clinical treatment centers. This self-report inventory is intended to determine how children between 8 and 12 years of age judge their own functioning in several specific domains and how they judge their global self-worth. The SPPC consists of six subscales, each consisting of six items. Five of the six subscales represent specific domains of self-concept: Scholastic Competence (SC), Social Acceptance (SA), Athletic Competence (AC), Physical Appearance (PA), and Behavioral Conduct (BC). The sixth scale measures Global Self-Worth (GS), which according to Harter (1985) is a more general concept. When a child fills out the SPPC, he or she first chooses which of the two statements applies to him or her and then indicates if the chosen statement is “sort of true for me” or “really true for me”. Scoring is done on a 4-point scale. The answer most indicative of a positive self-concept is scored “4”, and the answer least indicative of competence is scored “1”.

The psychometric properties of the five specific domain scales of the SPPC have been investigated using classical test theory and factor analytical approaches. Both research on the original English version and research on translated versions showed that a 5-factor model gave a reasonable fit. For example, Granleese and Joseph (1993) replicated the 5-factor structure obtained previously by Harter (1985) with American adolescents. Furthermore, they found a strong similarity in the correlations between subscale scores for girls and boys. Schumann et al. (1999) evaluated the SPPC in a biracial cohort for third graders and found a 5-factor solution for White girls, but not for Black girls. For Black girls, the physical appearance and the athletic competence domains were not yet fully differentiated. Thill et al. (2003) compared the SPPC with a questionnaire on actual behavior for children with and without spina bifida, and they found that the factor structure was similar for both groups.

For the Dutch version, Veerman et al. (2004) found a reasonable fit of the 5-factor model, where coefficient alpha for the subscales ranged from .68 (BC) to .83 (PA). In their study, van den Bergh and van Ranst (1998) also analyzed the Dutch version of the SPPC. They found that the factorial structure of the underlying self-concept was not exactly the same for fourth and sixth graders and that the SPPC was less reliable for boys than for girls and suggested that when performance of a
specific child has to be evaluated, the child is best situated in his or her gender group.

Recently, Meijer, Egberink, Emons, and Sijtsma (2008) used the SPPC to illustrate the usefulness of studying individual item score patterns. They identified children for whom the SPPC did not reflect their self-concept as a result of cognitive deficits. However, both the Meijer et al. (2008) study and the van den Bergh and van Ranst study (1998) suggested that the psychometric quality of the individual scales differed and that a more thorough psychometric analysis was needed to obtain a better picture of the characteristics of these scales.

The aim of the present study was to assess the psychometric quality of the SPPC scales by means of item response theory (IRT; Embretson & Reise, 2000) models. By using IRT modeling, a more detailed picture can be obtained about the functioning of individual items and scales and about the relation between trait scores (in this study self-concept scores) and item endorsement. Classical test theory and factor analytical approaches do not provide exhaustive item-level analysis. In the present study, we were in particular interested in (a) which items in each scale mainly determine the construct that is being measured and (b) whether the scales can reliably distinguish persons across different values of the latent trait scale. With IRT it is possible to investigate the relative contribution of each item to the measurement precision of the scale and to determine which items are most related to the construct being investigated. Thus, IRT can be used to obtain more refined information about the construct validity of the scales.

2.2 Method

2.2.1 Participants and Procedure

Part of the data that were reported in Meijer et al. (2008) were reanalyzed. Data were collected from 611 children between 8 and 12 years of age. The sample contained 343 girls (mean age = 10.19, $SD = 1.29$) and 268 boys (mean age = 10.17, $SD = 1.23$). Most children were White. These children came from five primary schools in the east of the Netherlands. Two schools are public primary schools and three are Catholic primary schools. Of the 611 children, there were 45 second graders (mean age = 8.33, $SD = 0.37$), 126 third graders (mean age = 8.73, $SD = 0.47$), 139 fourth graders (mean age = 9.79, $SD = 0.58$), 155 fifth graders (mean age = 10.73, $SD = 0.50$), and 146 sixth graders (mean age = 11.80, $SD = 0.55$). The research reported in
this study was part of a larger project in which information was obtained routinely from the children about their emotional and personal well-being.

2.2.2 Item Response Theory

IRT models are based on the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of items (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002). IRT explains the structure in the manifest responses by assuming the existence of a latent trait, denoted by the Greek letter θ. By means of IRT models, it is possible to locate a person’s θ and the characteristics of the items that make up the measurement instrument on the same metric (i.e., latent trait continuum).

In IRT, both nonparametric and parametric approaches can be distinguished. Nonparametric IRT models are based on less restrictive assumptions about the data and are, therefore, ideal instruments to explore the psychometric structure of tests and questionnaires. Parametric approaches are based on more restrictive assumptions but provide information that cannot be obtained using nonparametric approaches. In this study, we used Mokken’s nonparametric monotone homogeneity model (MMH; Sijtsma & Molenaar, 2002) to explore the psychometric structure of the SPPC and the parametric graded response model (GRM; Samejima, 1969, 1997) to obtain more detailed information about the measurement precision of the SPPC scales across the latent trait continuum. Furthermore, we used both approaches to obtain a detailed picture about the psychometric quality of the scales.

Mokken Scaling

The MMH is based on the assumptions of unidimensionality, local independence, and monotonicity. The model assumes that all items in a test or questionnaire measure the same latent trait (unidimensionality assumption), that a person’s response to one item is not influenced by the response to another item (local independence), and that the item response function is nondecreasing (monotonicity assumption). A more detailed description of these assumptions can be found in Sijtsma and Molenaar (2002) or Meijer and Banke (2004).

To check the assumptions of the MMH, several methods have been proposed. In this study, we used the coefficient $H_i$ for items and the coefficient $H$ for a set of items. Under the MMH, higher positive $H$ values reflect higher discrimination power of the items, and as a result, more confidence in the ordering of respondents by means of their total scores. This is also referred to as scalability, that is, the degree to
which a set of items are related to each other and form a scale. Items with high $H_i$ values discriminate well in the group in which they are used. $H_i$ values determine how well an item fits the scale. For practical test construction purposes, the following rules of thumb have been suggested. Weak scalability is obtained if $0.3 \leq H < 0.4$, medium scalability if $0.4 \leq H < 0.5$, and strong scalability if $0.5 \leq H < 1$ (Sijtsma & Molenaar, 2002). Values of $H$ smaller than 0.3 are considered evidence that the items are unscalable for practical purposes.

We used the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0; Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis for each subscale of the SPPC. We checked the assumptions of the MMH by inspecting the $H$ and $H_i$ coefficients. Because we suspected on the basis of the literature (van den Bergh & van Ranst, 1998) and on the basis of our own observations during test administration that there may be differences in model fit for boys and girls, we ran separate analyses for boys and girls.

**Graded Response Model**

To obtain a more detailed picture of the psychometric quality of the SPPC, we also analyzed the data with the GRM (Samejima, 1969, 1997). The GRM is suitable for analyzing ordered response categories, such as Likert-type rating scales. Several researchers used this model to analyze personality data, and there is a close relationship between the GRM and Mokken’s MMH model (Sijtsma & Molenaar, 2002). The MMH can be interpreted as a nonparametric version of the GRM, in the sense that both models assume unidimensional measurement, local independence, and nondecreasing item response functions.

The items in the GRM are defined by a discrimination parameter ($\alpha$; usually with numerical values between 0.5 and 2.5) and two or more location parameters ($\beta_m$; usually with numerical values between -2.5 and +2.5); the number of location parameters per item is equal to the number of response categories minus 1; thus, in our analysis, $4 - 1 = 3$. Like the $H_i$ coefficient, the magnitude of the discrimination parameter reflects the degree to which the item is related to the underlying latent trait. This means that for high $\alpha$ values the response categories accurately differentiate among trait levels. The location parameters reflect the spacing of the ordered response categories along the $\theta$ scale. The location parameter $\beta_m$ can be interpreted as the point at the latent trait continuum where there is a 50% chance of scoring in category $m$ or higher. Thus, respondents with a $\theta$ value higher than $\beta_m$ have more than 50% chance of responding in category $m$ or higher. These $\alpha$ and $\beta_m$ parameters are used to determine the probability of an examinee to respond in a
particular response category. These probabilities can be used to determine the category response functions, which describe the probability of responding in a particular response category conditional on $\theta$.

Figure 2.1 gives an example of the category response functions for two items of the BC scale of the SPPC for girls, Item 6 with a high estimated $\alpha$ value ($\hat{\alpha} = 2.38$; upper panel) and Item 1 with a low estimated $\alpha$ value ($\hat{\alpha} = 0.76$; lower panel). Moving from the lower to the higher end of the $\theta$ scale shows that first Category 1 is most likely (low $\theta$ levels), then Category 2, followed by Category 3, and, finally, Category 4 (high $\theta$ level). Furthermore, the middle category options are more peaked (higher $\hat{\alpha}$ values) for Item 6 than for Item 1.

![Category response functions for Item 6 and Item 1 of the Behavioral Conduct scale for girls.](image)

*Figure 2.1*: Category response functions for Item 6 of the Behavioral Conduct scale with $\hat{\alpha} = 2.38$ ($H_j = .39$; upper panel) and Item 1 of the Behavioral Conduct scale with $\hat{\alpha} = 0.76$ ($H_j = .22$; lower panel) for girls.
An important difference between Mokken scaling and the GRM is that in the former persons are assumed to have equal standard errors regardless of their position on the construct. In Mokken scaling, like in classical test theory, there is one reliability estimate. In parametric IRT, the concept of reliability is replaced by the concepts of item and test information. The standard error of a trait estimate is inversely related to the square root of the test information function. Thus, persons may have different standard errors depending on how discriminating a set of items is in different ranges of the latent trait. In general, items with larger discrimination parameters (i.e., the $\alpha$ parameters) provide relatively more information. The location parameters (i.e., the $\beta_m$ parameters) determine where the information is located. Item information is additive across the items administered and test information is maximized around the location parameters. Because information is inversely related to the standard error of measurement, this feature of IRT allows us to determine how precise a measure is for individuals in high, medium, and low trait ranges. We estimated the item parameters for the GRM using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003).

2.3 Results

2.3.1 Descriptive Statistics and Nonparametric Scaling

Table 2.1 depicts the mean item scores, item-test correlation, coefficient alpha, Guttman’s lambda-2, $H$, and $H_i$ coefficients for girls and boys. Like coefficient alpha, Guttman’s lambda-2 is a lower bound to test reliability, but it can be shown that Guttman’s lambda-2 is equal or greater than coefficient alpha and a more accurate lower bound than coefficient alpha. A first observation is that mean item scores are high, that is, most children have a positive self-concept on the respective subscales. Furthermore, we observe that there are differences in the psychometric quality of the different scales for both boys and girls and between boys and girls. This is reflected in the different mean values of the item-test correlations and $H$ values for the different scales.

In general, items are less scalable for boys than for girls. For boys, SC, SA, PA and BC form weak scales ($0.3 \leq H < 0.4$, although some items have $H_i < 0.3$; strictly

---

1 There were also some differences between the younger children and older children. For very young children, the items were less scalable than for older children. However, the very young children constitute only a small part of the sample. See Meijer et al. (2008) for more details.
Table 2.1
Descriptive Statistics for Girls and Boys.

<table>
<thead>
<tr>
<th>Item</th>
<th>Girls</th>
<th>Boys</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda^2$</td>
<td>$\alpha$</td>
<td>$M$</td>
</tr>
<tr>
<td>SC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.49</td>
<td>.43</td>
<td>.34</td>
</tr>
<tr>
<td>2</td>
<td>2.67</td>
<td>.45</td>
<td>.35</td>
</tr>
<tr>
<td>3</td>
<td>2.78</td>
<td>.56</td>
<td>.42</td>
</tr>
<tr>
<td>4</td>
<td>2.92</td>
<td>.50</td>
<td>.38</td>
</tr>
<tr>
<td>5</td>
<td>3.23</td>
<td>.54</td>
<td>.44</td>
</tr>
<tr>
<td>6</td>
<td>2.81</td>
<td>.59</td>
<td>.44</td>
</tr>
<tr>
<td>SA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3.20</td>
<td>.50</td>
<td>.38</td>
</tr>
<tr>
<td>8</td>
<td>3.23</td>
<td>.61</td>
<td>.45</td>
</tr>
<tr>
<td>9</td>
<td>2.77</td>
<td>.52</td>
<td>.40</td>
</tr>
<tr>
<td>10</td>
<td>3.27</td>
<td>.43</td>
<td>.33</td>
</tr>
<tr>
<td>11</td>
<td>2.86</td>
<td>.48</td>
<td>.37</td>
</tr>
<tr>
<td>12</td>
<td>3.18</td>
<td>.51</td>
<td>.39</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.98</td>
<td>.43</td>
<td>.30</td>
</tr>
<tr>
<td>14</td>
<td>2.83</td>
<td>.47</td>
<td>.33</td>
</tr>
<tr>
<td>15</td>
<td>3.10</td>
<td>.38</td>
<td>.27</td>
</tr>
<tr>
<td>16</td>
<td>2.62</td>
<td>.52</td>
<td>.38</td>
</tr>
<tr>
<td>17</td>
<td>3.41</td>
<td>.28</td>
<td>.22</td>
</tr>
<tr>
<td>18</td>
<td>2.99</td>
<td>.42</td>
<td>.31</td>
</tr>
<tr>
<td>PA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>3.26</td>
<td>.60</td>
<td>.50</td>
</tr>
<tr>
<td>20</td>
<td>3.05</td>
<td>.54</td>
<td>.46</td>
</tr>
<tr>
<td>21</td>
<td>3.22</td>
<td>.71</td>
<td>.57</td>
</tr>
<tr>
<td>22</td>
<td>3.17</td>
<td>.74</td>
<td>.59</td>
</tr>
<tr>
<td>23</td>
<td>3.21</td>
<td>.58</td>
<td>.48</td>
</tr>
<tr>
<td>24</td>
<td>3.30</td>
<td>.70</td>
<td>.58</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>2.87</td>
<td>.29</td>
<td>.22</td>
</tr>
<tr>
<td>26</td>
<td>3.28</td>
<td>.49</td>
<td>.36</td>
</tr>
<tr>
<td>27</td>
<td>3.19</td>
<td>.40</td>
<td>.29</td>
</tr>
<tr>
<td>28</td>
<td>2.93</td>
<td>.41</td>
<td>.30</td>
</tr>
<tr>
<td>29</td>
<td>2.93</td>
<td>.48</td>
<td>.35</td>
</tr>
<tr>
<td>30</td>
<td>3.21</td>
<td>.56</td>
<td>.39</td>
</tr>
<tr>
<td>GS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>3.23</td>
<td>.59</td>
<td>.48</td>
</tr>
<tr>
<td>32</td>
<td>3.35</td>
<td>.51</td>
<td>.41</td>
</tr>
<tr>
<td>33</td>
<td>3.50</td>
<td>.60</td>
<td>.47</td>
</tr>
<tr>
<td>34</td>
<td>3.47</td>
<td>.64</td>
<td>.50</td>
</tr>
<tr>
<td>35</td>
<td>3.39</td>
<td>.65</td>
<td>.50</td>
</tr>
<tr>
<td>36</td>
<td>3.14</td>
<td>.44</td>
<td>.37</td>
</tr>
</tbody>
</table>

Note. $M$ = mean item score; $r_{it}$ = item-test correlation; SC = Social Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth. $^a$ = Guttman’s lambda-2. $^b$ = coefficient alpha.
speaking only the SC scale forms a weak scale). The worst scale is the AC scale with \( H_i \) values between .17 and .26; thus boys are unscalable with respect to AC. For girls, in general, scalability coefficients are higher than for boys; for PA and GS they are much higher. These scales can be characterized as medium and strong scales. Intercorrelations between the total scores on the five basic scales ranged from \( r = .17 \) through \( r = .43 \) for boys and from \( r = .21 \) through \( r = .41 \) for girls. Correlations between the basic scales and GS ranged, however, from .24 (AC) to .68 (PA) for boys and between .33 (SC) and .76 (PA) for girls. At the end of the results section, we discuss IRT results when we combine items of the PA and GS scales.

### 2.3.2 Parametric IRT Analysis

In Table 2.2, we depicted the estimated item parameters for girls and boys. There are some interesting observations. A first observation is that there are items with extremely high \( \alpha \) values. For example, consider the PA scale for girls, here two items, Item 21 and Item 24, have \( \alpha \) values near 3.0, and a third item, Item 22, has an \( \alpha \) value of 3.55. One possibility is that this may point at item content redundancy, which is asking the same question twice. Items 21 and 22 use exactly the same phrasing but differ in one word: the Dutch word “lichaam” (English translation “body”) and the Dutch word “uiterlijk” (English translation “appearance”). So there is a subtle difference between the two items. Item 24 seems to be the item that is the shortest summary of PA (“I look good”).

Another possibility as suggested by Reise and Waller (2009) is that there is a highly skewed construct or quasi-trait. Many constructs in psychology, although assumed dimensional, are possibly, what they called “quasi-traits”, that is, traits that are only defined at one end of the latent trait scale. Reise and Waller mentioned constructs such as self-esteem, aggression, and spirituality. Self-perception as measured by the SPPC, which is conceptually related to self-esteem, also seems to be a quasi-trait. Consider the location parameters given in Table 2.2. For all scales the \( \beta_2 \) parameters are negative. Remember that persons with an estimated theta value, denoted by \( \hat{\theta} \), higher than \( \hat{\beta}_2 \) have more than 50% chance of responding in Category 2 or higher. Thus, persons with \( \hat{\theta} = 0 \), that is, with a mean score on self-concept, have more than 50% chance to answer in Category 3, and for the PA and the GS scale even in Category 4. This clearly indicates that the item locations are at the left side of the latent trait scale. One explanation may be due to the nature of the self-perception construct; items only differentiate between children with low self-perception because researchers are mainly interested in this end of the construct. A
Table 2.2

Estimated Item Parameters for Girls and Boys.

<table>
<thead>
<tr>
<th>Item</th>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.15</td>
<td>-1.78</td>
<td>0.01</td>
<td>1.70</td>
<td>1.28</td>
<td>-2.00</td>
<td>-0.56</td>
<td>0.88</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>-2.00</td>
<td>-0.16</td>
<td>1.06</td>
<td>1.48</td>
<td>-2.05</td>
<td>-0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>1.76</td>
<td>-1.73</td>
<td>-0.49</td>
<td>0.92</td>
<td>1.90</td>
<td>-1.50</td>
<td>-0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>1.55</td>
<td>-2.16</td>
<td>-0.70</td>
<td>0.70</td>
<td>1.46</td>
<td>-2.07</td>
<td>-0.72</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>1.91</td>
<td>-2.54</td>
<td>-1.32</td>
<td>0.31</td>
<td>1.69</td>
<td>-3.35</td>
<td>-1.60</td>
<td>0.32</td>
</tr>
<tr>
<td>6</td>
<td>1.98</td>
<td>-1.70</td>
<td>-0.54</td>
<td>0.86</td>
<td>2.16</td>
<td>-1.83</td>
<td>-0.61</td>
<td>0.51</td>
</tr>
<tr>
<td>SA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.51</td>
<td>-2.15</td>
<td>-1.17</td>
<td>0.01</td>
<td>1.87</td>
<td>-1.99</td>
<td>-1.33</td>
<td>-0.37</td>
</tr>
<tr>
<td>8</td>
<td>2.47</td>
<td>-1.66</td>
<td>-0.97</td>
<td>-0.12</td>
<td>3.28</td>
<td>-1.76</td>
<td>-1.16</td>
<td>-0.36</td>
</tr>
<tr>
<td>9</td>
<td>1.63</td>
<td>-1.32</td>
<td>-0.49</td>
<td>0.65</td>
<td>1.74</td>
<td>-1.34</td>
<td>-0.59</td>
<td>0.38</td>
</tr>
<tr>
<td>10</td>
<td>1.29</td>
<td>-2.92</td>
<td>-1.55</td>
<td>0.07</td>
<td>1.15</td>
<td>-3.04</td>
<td>-1.85</td>
<td>-0.08</td>
</tr>
<tr>
<td>11</td>
<td>1.44</td>
<td>-1.74</td>
<td>-0.72</td>
<td>0.70</td>
<td>1.27</td>
<td>-1.83</td>
<td>-0.74</td>
<td>0.57</td>
</tr>
<tr>
<td>12</td>
<td>1.52</td>
<td>-2.70</td>
<td>-1.39</td>
<td>0.43</td>
<td>0.96</td>
<td>-3.26</td>
<td>-2.04</td>
<td>0.54</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1.46</td>
<td>-1.80</td>
<td>-0.81</td>
<td>0.24</td>
<td>0.89</td>
<td>-2.20</td>
<td>-1.33</td>
<td>-0.11</td>
</tr>
<tr>
<td>14</td>
<td>1.70</td>
<td>-1.10</td>
<td>-0.46</td>
<td>0.17</td>
<td>0.79</td>
<td>-2.08</td>
<td>-1.00</td>
<td>0.08</td>
</tr>
<tr>
<td>15</td>
<td>1.18</td>
<td>-2.62</td>
<td>-1.30</td>
<td>0.40</td>
<td>1.51</td>
<td>-2.57</td>
<td>-1.39</td>
<td>0.09</td>
</tr>
<tr>
<td>16</td>
<td>2.02</td>
<td>-1.55</td>
<td>-0.22</td>
<td>1.20</td>
<td>1.20</td>
<td>-2.67</td>
<td>-0.97</td>
<td>1.18</td>
</tr>
<tr>
<td>17</td>
<td>0.81</td>
<td>-4.03</td>
<td>-2.36</td>
<td>-0.64</td>
<td>1.30</td>
<td>-2.50</td>
<td>-1.68</td>
<td>-0.68</td>
</tr>
<tr>
<td>18</td>
<td>1.25</td>
<td>-2.55</td>
<td>-1.10</td>
<td>0.79</td>
<td>1.52</td>
<td>-2.16</td>
<td>-0.98</td>
<td>0.54</td>
</tr>
<tr>
<td>PA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1.99</td>
<td>-2.07</td>
<td>-1.12</td>
<td>-0.02</td>
<td>1.63</td>
<td>-2.52</td>
<td>-1.77</td>
<td>-0.53</td>
</tr>
<tr>
<td>20</td>
<td>1.57</td>
<td>-1.62</td>
<td>-0.67</td>
<td>0.07</td>
<td>0.98</td>
<td>-2.59</td>
<td>-1.48</td>
<td>-0.63</td>
</tr>
<tr>
<td>21</td>
<td>2.81</td>
<td>-1.49</td>
<td>-0.75</td>
<td>-0.21</td>
<td>1.93</td>
<td>-1.86</td>
<td>-1.09</td>
<td>-0.49</td>
</tr>
<tr>
<td>22</td>
<td>3.55</td>
<td>-1.46</td>
<td>-0.65</td>
<td>-0.09</td>
<td>2.73</td>
<td>-1.78</td>
<td>-0.99</td>
<td>-0.42</td>
</tr>
<tr>
<td>23</td>
<td>2.15</td>
<td>-1.61</td>
<td>-0.87</td>
<td>-0.12</td>
<td>2.18</td>
<td>-1.86</td>
<td>-1.22</td>
<td>-0.54</td>
</tr>
<tr>
<td>24</td>
<td>2.99</td>
<td>-1.88</td>
<td>-1.09</td>
<td>0.01</td>
<td>2.31</td>
<td>-2.24</td>
<td>-1.52</td>
<td>-0.32</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.76</td>
<td>-3.08</td>
<td>-1.01</td>
<td>1.00</td>
<td>0.95</td>
<td>-2.47</td>
<td>-0.77</td>
<td>0.87</td>
</tr>
<tr>
<td>26</td>
<td>1.97</td>
<td>-2.62</td>
<td>-1.45</td>
<td>0.19</td>
<td>1.74</td>
<td>-2.43</td>
<td>-1.16</td>
<td>0.54</td>
</tr>
<tr>
<td>27</td>
<td>1.25</td>
<td>-3.26</td>
<td>-1.63</td>
<td>0.46</td>
<td>1.00</td>
<td>-3.22</td>
<td>-1.75</td>
<td>0.49</td>
</tr>
<tr>
<td>28</td>
<td>1.09</td>
<td>-2.69</td>
<td>-0.74</td>
<td>0.61</td>
<td>1.16</td>
<td>-2.37</td>
<td>-0.49</td>
<td>0.72</td>
</tr>
<tr>
<td>29</td>
<td>1.73</td>
<td>-2.26</td>
<td>-0.73</td>
<td>0.68</td>
<td>1.79</td>
<td>-1.65</td>
<td>-0.30</td>
<td>0.91</td>
</tr>
<tr>
<td>30</td>
<td>2.38</td>
<td>-2.22</td>
<td>-1.31</td>
<td>0.33</td>
<td>2.58</td>
<td>-2.17</td>
<td>-1.09</td>
<td>0.55</td>
</tr>
<tr>
<td>GS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1.87</td>
<td>-2.31</td>
<td>-1.19</td>
<td>0.17</td>
<td>0.91</td>
<td>-3.85</td>
<td>-2.71</td>
<td>-0.39</td>
</tr>
<tr>
<td>32</td>
<td>1.49</td>
<td>-2.22</td>
<td>-1.41</td>
<td>-0.37</td>
<td>1.46</td>
<td>-2.12</td>
<td>-1.49</td>
<td>-0.64</td>
</tr>
<tr>
<td>33</td>
<td>2.53</td>
<td>-2.28</td>
<td>-1.42</td>
<td>-0.40</td>
<td>2.03</td>
<td>-2.61</td>
<td>-2.10</td>
<td>-0.79</td>
</tr>
<tr>
<td>34</td>
<td>2.94</td>
<td>-2.06</td>
<td>-1.26</td>
<td>-0.41</td>
<td>3.05</td>
<td>-1.76</td>
<td>-1.39</td>
<td>-0.70</td>
</tr>
<tr>
<td>35</td>
<td>3.19</td>
<td>-1.84</td>
<td>-1.06</td>
<td>-0.32</td>
<td>2.40</td>
<td>-2.08</td>
<td>-1.57</td>
<td>-0.70</td>
</tr>
<tr>
<td>36</td>
<td>1.21</td>
<td>-2.82</td>
<td>-1.25</td>
<td>0.33</td>
<td>0.89</td>
<td>-3.69</td>
<td>-1.86</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Note. SC = Social Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.
high level of self-perception (i.e., high self-concept) is healthy, whereas a low level of self-perception (i.e., low self-concept) may be problematic. Therefore, the aim of measuring self-perception is to detect children with low self-perception, and therefore, items are written that measure low self-perception. An alternative interpretation is that most people in today’s Western world have high self-esteem (see Baumeister, 1993).

Another interesting observation concerns the SA and the GS scales for boys. For the SA scale, Item 8 (“Having many friends”) has a very high discrimination parameter ($\hat{\alpha} > 3$), whereas Items 10, 11, and 12 have low discrimination parameters (around $\hat{\alpha} = 1$). Conceptually this means that Item 8 defines SA, whereas Items 10, 11, and 12 are much less related to SA. For example, Item 10, “Doing many things alone or with others”, may indicate social acceptance but may also be related to a preferred way of doing things. It is clear that, in general, high $\alpha$ values go together with high $H_i$ as expected. In this data set, $H_i \geq .30$ corresponds to $\hat{\alpha} \geq 1.20$ and $H_i \geq .40$ correspond to $\hat{\alpha} \geq 1.50$. Although $H_i$ coefficients and $\alpha$ parameters both indicate the slope of the item response function, both statistics are also sensitive to different characteristics of the data. $H_i$ is strongly influenced by the probability distribution of the latent trait values, which also prevents a researcher from using a scale in a population where it cannot discriminate between persons. In the literature there is a strong emphasize on selecting items with $H_i$ values larger than some lower bound as, say, $H_i = .3$. We observe, however, that a researcher should also be careful when $H_i$ values are very high. We agree with Sijtsma and Molenaar (2002) that “one should find measurement instruments that measure one meaningful psychological ability or trait at a time” (p. 19). The question here is what is “meaningful”. As we discussed, repeating items with a similar content will result in scales with high $H$ values but, sometimes, extremely small-band constructs. It is, therefore, very important to inspect the content of the items and the scales and, perhaps most important, how the content of the items is interpreted. Especially for special groups such as young children or clinical patients, items may be interpreted differently than a researcher is suspecting.

Because large $H_i$ values and large $\alpha$ parameters go together, strong Mokken scales may also be the result of violations of local independence and the result of narrow-band constructs. Although these scales are very reliable, one should be careful to include items in a scale that are not semantically similar. High $H_i$ values may also point at items that define the construct (“I am often depressed” in a depression list). For example, Emons, Meijer, and Denollet (2007) analyzed Negative Affectivity items and found that the dysphoria item “is often down in the dumps” had an $\hat{\alpha}$
parameter value of 3.61, whereas an item such as “is easily irritated” had an \( \hat{\alpha} \) parameter value of 1.45.

### 2.3.3 Measurement Precision

Figure 2.2 displays the test information functions for the six subscales of the SPPC for girls (for boys similar results were obtained). For all scales the highest information is located at the lower trait values, that is, between scale scores \( \hat{\theta} = -2 \) and 0. This can be explained by noting that most item locations are situated at the lower trait ranges, that is, all items are relatively easy or popular. Remember that information is inversely related to the standard error of measurement and this feature thus allows us to determine the measurement precision of the SPPC scales for children in low, medium, and high trait ranges.

From Figure 2.2, it is clear that for girls the scales do not provide precise measurement across the whole scale and that even in the parts of the scale where there is some measurement precision (with the exception of the PA and the GS scales) broad confidence intervals result. For example, the six items that together measure AC provide information of around 4 for values between \( \hat{\theta} = -2 \) and 0, which corresponds to \( SE = 0.50 \). Thus, for a child with \( \hat{\theta} = -1 \), the 95% confidence interval is between -2 and 0, which is clearly too broad to base any substantive conclusion on. Even for the subscales that together form a weak Mokken scale, such as SC and SA, using only six items often result in very broad confidence intervals. For example, for SC and scale score \( \hat{\theta} = 0 \) information equals 5, which corresponds to \( SE = .45 \) and thus a 95% confidence band between -.90 and +.90. Similar results were obtained for boys. Thus, the take-home message is that these subscales only provide some measurement precision at the left part of the \( \theta \) scale (i.e., for total scores smaller than 15, which corresponds to \( \hat{\theta} = 0 \)) and that even these score profiles should be interpreted very carefully because for most scales broad confidence bands exist.
Chapter 2

2.3.4 Combined Scales

Because correlations between the PA and GS scales were high, we investigated whether items of these scales cluster together. Both for boys and for girls this was the case. Consider the item parameters and $H_i$ values in Table 2.3 for the combined scale (PA and GS together) for girls. For most items, the $H_i$ values of the combined scale are comparable with the $H_i$ values for the individual scales (depicted in Table 2.1). For 7 out of the 12 items, the $\hat{\alpha}$ parameters for the combined scale were higher than for the individual scales (Table 2.2). This clearly indicates item redundancy, because it is expected that when we measure a broader trait as a result of combining two scales, item discrimination will decrease.
Table 2.3

*Estimated Item Parameters and H, Coefficients for 12 items of the GS and PA Scale as One Scale for Girls (H = .48, When Item 32 Is Removed, H = .50).*

<table>
<thead>
<tr>
<th>Item</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$H_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA 19</td>
<td>2.13</td>
<td>-1.79</td>
<td>-0.89</td>
<td>0.14</td>
<td>.48</td>
</tr>
<tr>
<td>PA 20</td>
<td>1.55</td>
<td>-1.43</td>
<td>-0.51</td>
<td>0.22</td>
<td>.41</td>
</tr>
<tr>
<td>PA 21</td>
<td>2.60</td>
<td>-1.31</td>
<td>-0.59</td>
<td>-0.05</td>
<td>.51</td>
</tr>
<tr>
<td>PA 22</td>
<td>3.65</td>
<td>-1.23</td>
<td>-0.48</td>
<td>0.06</td>
<td>.57</td>
</tr>
<tr>
<td>PA 23</td>
<td>2.23</td>
<td>-1.38</td>
<td>-0.69</td>
<td>0.03</td>
<td>.45</td>
</tr>
<tr>
<td>PA 24</td>
<td>3.32</td>
<td>-1.60</td>
<td>-0.86</td>
<td>0.16</td>
<td>.55</td>
</tr>
<tr>
<td>GS 31</td>
<td>1.76</td>
<td>-2.16</td>
<td>-1.03</td>
<td>0.33</td>
<td>.45</td>
</tr>
<tr>
<td>GS 32</td>
<td>1.25</td>
<td>-2.27</td>
<td>-1.36</td>
<td>-0.22</td>
<td>.36</td>
</tr>
<tr>
<td>GS 33</td>
<td>2.59</td>
<td>-2.03</td>
<td>-1.19</td>
<td>-0.21</td>
<td>.51</td>
</tr>
<tr>
<td>GS 34</td>
<td>2.69</td>
<td>-1.88</td>
<td>-1.07</td>
<td>-0.24</td>
<td>.51</td>
</tr>
<tr>
<td>GS 35</td>
<td>4.14</td>
<td>-1.50</td>
<td>-0.81</td>
<td>-0.12</td>
<td>.57</td>
</tr>
<tr>
<td>GS 36</td>
<td>1.24</td>
<td>-2.55</td>
<td>-1.04</td>
<td>0.49</td>
<td>.35</td>
</tr>
</tbody>
</table>

*Note.* PA = Physical Appearance; GS = Global Self-Worth.

Another observation is that half of the amount of information is due to three items. In Figure 2.3, we provide the test information curve for all 12 items in the combined scale. Test information equals 23 between $\hat{\theta} = -2$ and 0. When we select Item 22 (“Satisfied with how I look like”) and Item 24 (“Happy with how I look like”) of the PA scale, and Item 35 of the GS scale (“Happy with who I am”), we obtain test information of 12. These three items provide half of the information of a scale that is 4 times as long. Thus, the combination of GS items and PA items is for a large part a “Happy with how I look like” scale.

We further conclude that, although some authors echo Harter’s (1985) idea that Global Self-Worth “is not a measure of general competence and as such should be considered an independent component of the scale” (e.g., Shevlin, Adamson, & Collins, 2003, p. 1995), we found high correlations between PA and GS. This was also found by Schumann et al. (1999) for girls. GS is thus not an independent component of the scale, but is heavily saturated with PA. For children, physical appearance heavily determines global self-worth. This close association between satisfaction with physical appearance and global self-worth has been hypothesized to contribute to the risk for the development of eating disorders in adolescent girls and may result from an overemphasis by the popular media on the importance of looking good.
Chapter 2

2.4 Discussion

What did we learn from our analyses? Our analyses suggest that when interpreting SPPC scores, care should be taken in interpreting the different subscale scores. First, measurement precision differs across scales and across the latent trait values within a scale. Like in many typical performance measures (see Reise & Waller, 2009), self-perception is only defined at one end of the scale, that is, self-perception in a non-clinical population is not a bipolar trait. The trait is only defined at the lower end of the trait scale, where children are situated with a relatively low self-concept. This implies that children with a medium to high self-concept cannot be distinguished from each other using the SPPC.

Furthermore, although each scale of the SPPC consists of 6 items, some constructs can be measured by 4 items (the PA scale), whereas other constructs need more items (the SC scale). There is, however, a catch. For the SPPC, it is clear that for the PA and GS scales there are only three items needed to obtain a scale with information around 12, but these items are semantically similar or have the same meaning for young children. Our example showed that Items 22 and 24 are semantically similar and that although Item 35 of the GS scale ("Happy with who I am") seems to have a different content, additional qualitative research showed that this item is interpreted by many children as "Happy with how I look like".

![Figure 2.3: Test information function for the combined scale (Physical Appearance and Global Self-Worth scales together) for girls.](image)
But scales should not be very short. First, when we use measurement precision as a criterion we select items that basically repeat the same question. Therefore, we distrust personality scales that consist of only a few items. Although we often encounter phrases like “given a standard error of 0.35, only two to three items are needed to evaluate a single dimension”, we think that a researcher should be extremely careful when scaling persons on the basis of two or three items. Consider, for example, the subscale SA for boys. We found (roughly) the following item information values for $\theta$ values in the range (-2, 0) for the six items (1, 3, 1, 0.4, 0.4, and 0.2). This implies that if a researcher accepts a standard error of 0.5, only one or two items are needed. However, good measurement is not only about precise measurement of the latent trait, it is also about replication and validity (in fact it is all about replication and validity).

What is often neglected is that very few items will result in an unacceptably low classification consistency (see e.g., Emons, Sijtsma, & Meijer, 2007). Classification consistency refers to the percentage of persons assigned to the same diagnostic category (such as low self-concept) by two hypothetical independent replications of the test. Emons, Sijtsma, & Meijer (2007) showed that for short tests (between 6 and 12 items) classification into two categories (e.g., treatment and nontreatment groups) resulted in at most 50% correct classifications, whereas results for longer tests (20 and 40 items) were much better. Although one may argue that short measures will never be used in practice to classify a person, for example, for a treatment, we noticed that especially in medicine and health psychology there is a trend to use extremely short scales, for example, through the Internet to classify persons as depressed or anxious (e.g., Taylor & Deane, 2002).

Thus, there is a dilemma when using inventories in personality measurement. On the one hand, we do not want to administer long self-reports that consist of items that basically repeat the same question over and over; on the other hand, inventories cannot be too short because results are difficult to replicate. What to do? A way out of this dilemma is perhaps the observation that in practice questionnaires are seldom used as the only indicator to classify a person into a category. Often interviews, observation, and other tests and questionnaires are being used. Combining information from different (sometimes unreliable) sources will result in acceptable replication rates and acceptable validity. Cronbach (1954) already showed that, given a fixed testing time, it is often better for personnel selection or treatment referral to use many short tests covering many dimensions than one (large) accurate test. This, thus, justifies the use of the SPPC as a general multidimensional measure of self-perception. At the same time, it warns the psychologist to overemphasize the
meaning of the total scores on the individual scales and thus - perhaps more important - subtest profiles, when they are not combined with other information. Furthermore, because researchers are often interested in measuring broad-band constructs it seems a good strategy, when constructing typical performance measures, to strive for a number of weak or medium Mokken scales. The alternative strategy to strive for strong Mokken scales consisting of many items will result in scales with items with similar content.

Because self-perception is a quasi-trait in a community sample, it has consequences for longitudinal research. For example, Shapka and Keating (2005) studied the longitudinal changes in multiple domains of self-concept over a 2-year period in adolescence. They found that most domains of self-concept increase with age, although perceived scholastic competence decreased. When SPPC scales are used in a longitudinal setting to measure change over time, it is very important to realize that the SPPC provides different measurement precision across the trait range. Clearly, the SPPC scales are a lot more sensitive to change in the lower regions of the latent trait because standard errors are much smaller there than in the higher regions of the scale.

A limitation of this study was the relatively small sample size for boys and girls for estimating the item parameters in the GRM analyses. Although we recognize this, we observe that the results of the GRM analysis were confirmed by the Mokken scale analysis for which less persons are needed. Also, our aim was not to provide definitive parameters estimates but rather to address specific questions regarding the strengths and weaknesses of the SPPC. Finally, in this study data were obtained from a nonclinical population. Because the SPPC is intended to be used in both clinical and nonclinical populations, future research may investigate the psychometric quality of the SPPC in a clinical population.

2.5 Recommendations

Given our perspective on the use of the SPPC in practice, we have the following recommendations and take-home messages:

(1) Psychologists and researchers should realize that contrary to earlier research, the subscales GS and PA are measuring similar concepts. Thus, the subscale GS should not be interpreted as an overarching scale, that is, as a scale that measures “general self-concept”. Because of its high correlation with PA, it seems more of an appearance scale.
(2) Researchers should be very careful to use the total scores on the different subscales to distinguish children with a medium to high self-concept. Measurement in these ranges of the latent trait is very unreliable. Although, some subscales (PA) are suited to distinguish children with a low self-concept from children with a medium to high self-concept, in general the scales of the SPPC do not provide very accurate latent trait estimates. PA is a strong scale, whereas AC is a very weak scale. We, therefore, suggest that future research may reconsider a revision of several subscales of the SPPC. Using both content information as well as information from IRT analysis, items can be selected that allow for constructs that are broad enough to have any empirical validity and that provide acceptable measurement precision to distinguish children on the different important constructs of self-concept. We realize that it will not always be easy (perhaps sometimes impossible) to come up with subscales that are both reliable and that measure constructs that are broad enough for prediction, but we think that IRT can be of help in obtaining better SPPC scales than the existing ones.
Chapter 3

Detection and Validation of Unscalable Item Score Patterns Using Item Response Theory: An Illustration with Harter’s Self Perception Profile for Children

Abstract

We illustrate the usefulness of person-fit methodology for personality assessment. For this purpose, we use person-fit methods from item response theory. First, we give a nontechnical introduction to existing person-fit statistics. Second, we analyze data from Harter’s (1985) Self-Perception Profile for Children in a sample of children ranging from 8 to 12 years of age (N = 611) and argue that for some children the scale scores should be interpreted with care and caution. Combined information from person-fit indices and from observation, interviews, and self-concept theory showed that similar score profiles may have a different interpretation. For some children in the sample, item scores did not adequately reflect their trait level. Based on teacher interviews, this was found to be due most likely to a less developed self-concept and/or problems understanding the meaning of the questions. We recommend investigating the scalability of score patterns when using self-report inventories to help the researcher interpret respondents’ behavior correctly.

This chapter has been published as:
3.1 Introduction

*A psychologist trying to understand an individual’s personality is a bit like a detective trying to solve a mystery: clues may abound, but the trick is to correctly interpret what they mean.* (Funder, 2004, p. 20)

There exists a tradition in personality assessment to detect invalid test scores using different types of validity scales such as the Variable Response Inconsistency Scale and the True Response Inconsistency Scale of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 2001), although the usefulness of these scales is not undisputed (e.g., Piedmont, McCrae, Riemann, & Angleitner, 2000). In the psychometric and personality literature (e.g., Meijer & Sijtsma, 2001; Reise & Waller, 1993), it has been suggested that invalid test scores can also be identified through studying the configuration of individual item scores by means of person-fit statistics that are proposed in the context of item response theory (IRT; Embretson & Reise, 2000). On the basis of an IRT model, observed and expected item scores can be compared, and many unexpected item scores alert the researcher that the total score may not adequately reflect the trait being measured.

The literature on person fit is mainly technical in the sense that there are many studies that have been devoted to the psychometric characteristics of the statistics and tests (such as the correct sampling distribution), but there are very few studies that have illustrated the usefulness of these statistics in practice (e.g., Meijer & Sijtsma, 2001). There is a gap between the often very sophisticated articles devoted to the psychometric characteristics of several statistical tests and measures on the one hand and the articles that describe the practical usefulness of these measures on the other hand. Rudner, Bracey, and Skaggs (1996) remarked that “In general, we need more clinical, practically oriented studies that find aberrant patterns of responses and then follow up with the respondents. We know of no studies that empirically investigate what these respondents are like. Can anything meaningful be said about them beyond the fact that they do not look like typical respondents?” (p. 107).

In this study, we tried to integrate psychometric analysis with information from qualitative sources to make judgments about the validity of an individual’s test score. More specifically, the aim of this study was to (a) explore the usefulness of person-fit statistics to identify invalid test scores using real data and (b) validate information obtained from IRT using personality theory and qualitative data obtained from observation and interviews.
This article is organized as follows. First, we explain the usefulness of IRT to investigate the quality of individual item score patterns. Second, we provide a nontechnical background to person-fit analysis in the context of nonparametric IRT. Finally, we illustrate the practical usefulness of person-fit statistics in the context of personality assessment using the Self-Perception Profile for Children (SPPC; Harter, 1985).

3.2 Item Response Theory and Individual Score Patterns

3.2.1 IRT Measurement Model

The use of IRT models in the personality domain is rapidly becoming more popular. Although empirical similarities of tests and inventories constructed according to classical test theory and IRT do exist, IRT offers better solutions to many psychometric problems than classical test theory (e.g., Embretson & Reise, 2000). The detection of invalid test scores is an interesting and practically useful example (Meijer, 2003). Reise and Henson (2003) provided other examples of the usefulness of IRT models to analyze personality data.

In most IRT models, test responses are assumed to be influenced by a single latent trait, denoted by the Greek letter $\theta$. For dichotomous (true, false) data, the goal of fitting an IRT model is to identify an item response function (IRF) that describes the relation between $\theta$ and the probability of item endorsement. In most IRT models, it is assumed that the probability of item endorsement should increase as the trait level increases; thus, IRFs are monotonically increasing functions (for an exception, see e.g., Ramsay, 1991). In Figure 3.1 examples are given of several IRFs.

Figure 3.1: Example item response functions for different models.
More formally, the IRF, denoted $P_i(\theta)$, gives the probability of endorsing an item $i$ ($i = 1, \ldots, k$) as a function of $\theta$. It is the probability of a positive response (i.e., "agree" or "true" in response to a positively phrased statement about one’s personality) among persons with the latent trait value $\theta$. For dichotomous items, $P_i(\theta)$ often is specified using the one-, two-, or three-parameter logistic model (1PLM, 2PLM, 3PLM; see Embretson & Reise, 2000). These models are characterized by an S-shaped IRF. Examples are the IRFs indexed 1, 2, and 3 in Figure 3.1.

**Nonparametric IRT**

In this study, we used the nonparametric Mokken model of monotone homogeneity (MMH; e.g., Sijtsma & Molenaar, 2002). This model assumes that the IRFs are monotonically increasing, but a particular shape for the IRF is not specified. Thus, in Figure 3.1, all IRFs can be described by the MMH. Notice that the IRFs of Items 4 and 5 are not S-shaped and thus cannot be described by a logistic model. Nonparametric models have the advantage that they are more flexible than parametric models and therefore sometimes better suited to describe personality data than parametric models (see Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001, and Meijer & Baneke, 2004, for an extensive discussion in the personality domain, and Junker & Sijtsma, 2001, for a discussion of similarities and differences between parametric and nonparametric models). Another advantage is that the MMH is a relatively simple model that is easy to communicate to applied researchers.

The MMH allows the ordering of persons with respect to $\theta$ using the unit weighted sum of item scores (i.e., total raw score). Using the MMH, we first investigate if the model fits the data before we use the total score to rank order persons. Investigating the fit of the model has the advantage that items can be identified that do not contribute (or that contribute even negatively) to the rank ordering of persons.

The MMH is a probabilistic approach to the analysis of item scores that replaces the well-known Guttman (1950) model. According to the Guttman model, it is not allowed that a person endorses a less popular item while rejecting a more popular item. For example, when measuring depressed suicidal ideation, every person that endorses the statement “I have recently considered killing myself” is expected to also endorse the statement “I don’t seem to care what happens to me” (relative to the previous item, this item is less extreme or, in a terminology that is sometimes used, more popular). However, in practice, when analyzing personality data, many “errors”
are found against the Guttman model. That is, several persons do not endorse the “don’t care what happens” statement but they do endorse the “killing myself” statement. Thus, perfect scales (Guttman scales) consisting of perfect item pairs that do not induce score reversals as described in the example rarely exist in practice. The ideal, as usual, can only be approximated. Obviously, the Guttman model is based on an unrealistic requirement of response behavior. Probabilistic models such as the Mokken model allow deviations (i.e., errors from the perspective of the Guttman model) from this requirement within certain limits defined by the specific probabilistic model.

As for the MMH for dichotomous items, the MMH for polytomous items that we use in this study assumes increasing response functions. The only difference is that the assumption is now applied to the so-called item step response function (ISRF). An item step is the imaginary threshold between adjacent ordered response categories. As an example, imagine the positively worded personality item “I have strange and peculiar thoughts” having three ordered answer categories (disagree, agree, strongly agree). It is assumed that the participant first ascertains whether he or she agrees enough with the statement to take the first item step (from disagree to agree). If not, the first item step equals 0, and the item score also equals 0. If the answer is affirmative, the item step equals 1, and the participant has to ascertain whether the second step (from agree to strongly agree) can be taken. If not, the second item step equals 0, and the item score equals 1. If the answer is affirmative, the second item step score equals 1, and the item score equals 2. The ISRF describes the relation between the probability that the item step score equals 1 and $\theta$. An item with three ordered answer categories has two item steps and consequently, two ISRFs, one for each item step. The MMH assumes that each of the ISRFs is monotone nondecreasing in $\theta$. Nondecreasingness of ISRFs can be investigated by inspection of the regression of the observed item step score on the rest score. The rest score is defined as the score on the other $k$-1 items without the score on the item $i$. The ISRF should be a monotonely nondecreasing function of the rest score.

In Figure 3.2, examples of the four ISRFs of a five-category item are given that are in concordance with the MMH. As with the MMH for dichotomous items, measurement by means of the MMH for polytomous items also uses the total (or rest) score for ordering respondents on $\theta$.

To check whether the ISRFs are monotonically nondecreasing, several methods have been proposed. In this study, we used the coefficient $H_i$ for items ($i = 1, \ldots, k$) and coefficient $H$ for a set of items. Increasing values of $H$ and $H_i$ between .3 and 1 (maximum) mean that the evidence for monotone increasing ISRFs is more
convincing, whereas values below .3 indicate violations of increasing ISRFs (for a discussion of these measures, see e.g., Meijer & Banke, 2004, or Sijtsma & Molenaar, 2002). Furthermore, weak scalability is obtained if \( .3 \leq H < .4 \), medium scalability if \( .4 \leq H < .5 \), and strong scalability if \( .5 \leq H < 1 \) (Sijtsma & Molenaar, 2002, pp. 60-61). Values of \( H \) smaller than .3 are considered evidence that the item set is unscalable for practical purposes.

![Figure 3.2: Example of item step response functions. Note. Rest score = Total score – Score on this item; prop. = proportion. The horizontal axis shows groups of adjacent rest scores that were joined due to low frequencies for individual rest scores.](image)

3.2.2 Studying Individual Item Score Patterns

By means of an IRT model, it is possible to predict a person’s answering behavior when confronted with a particular set of questionnaire items. Let us illustrate this by means of Figure 3.3. For the sake of simplicity, we depicted five IRFs that do not intersect across the latent trait range. These IRFs comply with the Rasch (1960) model. Assume that we have an estimate of someone’s trait level to be \( \hat{\theta} = 0 \), then the probability of endorsing Item 1 equals .9 (this is the most popular item), and the probability of endorsing Item 5 equals .1 (this is the least popular item). Suppose now that the items are ordered from most popular to least popular and that a person endorses three items; then the item score pattern that has the highest probability of occurrence is 11100, and the item score pattern with the lowest probability of
occurrence is 00111. This second pattern thus is unexpected, and it may be questioned whether the total score of 3 has the same meaning for both patterns.

![Figure 3.3: Five item response functions following the Rasch model.](image)

**Person-fit Statistics**

Several indices and statistical tests have been proposed to identify unexpected item score patterns (Meijer & Sijtsma, 2001). A very simple person-fit statistic is the number of Guttman (1950) errors found in an individual’s item score pattern. Given that the items are ordered according to decreasing level of popularity, for dichotomous item scores, the number of Guttman errors is simply defined by the number of 0 scores to the left of each 1 score. Thus, for example, the pattern (110101) contains three Guttman errors [three (0,1) item pairs]. This index was also used by Meijer (1994) and Emons, Sijtsma, and Meijer (2005) and was found to be one of the best performing person-fit indices.

For polytomous items, the popularity of the item steps can be determined, and the item steps can then be ordered according to decreasing popularity. A Guttman error consists of endorsing a so-called less popular item step in one item while not endorsing a more popular item step in another item. To illustrate this, consider a scale that consists of six items with four response alternatives (coded 1-4). This implies that there are three item steps per item (from 1-2, 2-3, and 3-4). Thus there are $6 \times 3 = 18$ item steps for each person. As an example, we consider two score patterns that we found in the SPPC we will analyze shortly. One observed score
pattern is 323322. From this pattern, we obtained the item step score pattern 1111001101000000, with item steps ordered to decreasing popularity. These popularities were obtained from the observed proportion of respondents who endorse a particular response category or a higher category. This pattern contains 7 Guttman errors (7 pairs of 0 before 1 scores). The other pattern is 144414, with the corresponding item step score pattern 01110011010101111, which contains 45 Guttman errors. Further details on how Guttman errors are counted for polytomous items can be found in Sijtsma and Molenaar (2002).

**Person-fit Studies in the Personality Domain**

Studies in personality context used simulated data to investigate whether person-fit statistics could be used as alternatives to social desirability and lying scales to identify dishonest respondents. Results have been mixed. Zickar and Drasgow (1996) concluded that person-fit statistics were useful alternatives to validity scales, whereas Ferrando and Chico (2001) found that person-fit statistics were less powerful than validity scales. In one of the few studies using empirical data, Reise and Waller (1993) investigated whether person-fit statistics may help to identify persons who do not fit a particular conception of a personality trait (called “taledness”; Tellegen, 1988). Reise and Waller investigated the usefulness of a person-fit statistic to investigate traitsedness by analyzing 11 unidimensional personality subscales. Results showed that person-fit statistics could be used to explore the fit of an individual's response behavior to a personality construct. Reise and Waller used scales that had well-defined content clusters and were therefore able to link unscalability to aberrant responding for specific types of item content. In general, however, one should be careful when interpreting misfitting item score patterns. Reise and Waller discussed that interpreting misfitting item score patterns as an indicator of traitsedness variation is difficult. Possible causes may be response faultliness, misreading, or random responding. Thus, person-fit statistics allow the identification of a deviant item score pattern but not the recovery of the mechanism that created a deviant item score pattern.

As the personality researcher usually does not know the cause of an atypical item score pattern, for a better understanding of the potential causes, background information about individual persons needs to be incorporated into the diagnostic process. Depending on the application, such information may come from previous psychological-ability and achievement testing, school performance (tests and teacher’s accounts), clinical and health sources (e.g., about dyslexia, learning and memory problems) or social-economic indicators (e.g., related to language problems.
at home). In this study, we combined information from person-fit statistics with auxiliary information from personality theory and a person’s personal history. Although in many studies it has been suggested that quantitative and qualitative information should be combined, there are few studies in which this has been done (for an example, see Emons, 2003, chap. 6).

3.3 Method

3.3.1 Instrument
Data were analyzed from the official Dutch translation of Harter’s (1985) SPPC (Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004). This self-report inventory is intended to determine how children between 8 and 12 years of age judge their own functioning in several specific domains and how they judge their global self-worth. The SPPC consists of six subscales each consisting of six items. Five of the six subscales represent specific domains of self-concept: Scholastic Competence (SC), Social Acceptance (SA), Athletic Competence (AC), Physical Appearance (PA), and Behavioral Conduct (BC). The sixth scale measures Global Self-Worth (GS), which is a more general concept. When a child fills out the SPPC, he or she first chooses which of the two statements applied to him or her and then indicates if the chosen statement is “sort of true for me” or “really true for me”. Scoring is done on a 4-point scale ranging from “1” to “4”. The answer most indicative of competence is scored 4, and the answer least indicative of competence gets a score 1.

To date, psychometric properties (multidimensional structure, invariance across groups) of the SPPC have been investigated mainly using classical test theory and factor analytical approaches. Veerman et al. (2004, pp. 21- 25) showed that for the Dutch version of the SPPC, a five-factor model gave a reasonable fit. Coefficient alpha for the subscales ranged from .68 (BC) to .83 (PA). In their study, van den Bergh and van Ranst (1998) also analyzed the Dutch version of the SPPC. They found that the factorial structure of the underlying self-concept was not exactly the same for fourth and sixth graders and that the SPPC was less reliable for boys than for girls. They suggested that when performance of a specific child has to be evaluated, the child is best situated in his or her gender and age groups.

3.3.2 Participants and Procedure
Data were collected from 702 children between 7 and 13 years of age. The sample contained 391 girls and 311 boys (mean age = 9.82). Most children were White.
These children came from five primary schools in the east of the Netherlands. From this dataset, we removed 91 children younger than 8 years of age because they did not belong to the population for which the SPPC is intended and were too young to fill out the SPPC adequately. Five children were older than 12 years, but they were not removed from the data because they could fill out the SPPC adequately. The final sample consisted of 611 children, of which 343 were girls and 268 boys (mean age = 10.18). The research reported in this article was part of a larger project in which information was obtained routinely from the children about their emotional and personal well-being.

The inventory was administered to children in classroom groups. Before the children completed the scales, we provided standardized oral instructions. During the test administration, I. J. L. Egberink was available for further clarification. After test administration, we investigated whether the MMH fitted the data (model fit), constructed a profile of the subtest scores for each child, and computed the person-fit statistics. The results were discussed with the teacher. During these conversations, we explained the subtest profiles and also discussed possible explanations for inconsistent response behavior for those children that were flagged as inconsistent by the person-fit statistic.

Four of the five schools that participated in this study allowed us to readminister the SPPC for children showing inconsistent response behavior. This enabled us to determine the stability of the results. Before we readministered the SPPC, we explained to the children that we randomly picked some children for research purposes. We did not explain that they produced irregular patterns because we wanted to keep retest conditions as similar as possible as the first time, and upsetting children might easily cause these conditions to be different. The main difference was that at the second administration, the children were tested in smaller groups than at the first administration.

### 3.3.3 Analysis

**Model Fit**

We used the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0; Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis for each subscale of the SPPC and to compute the fit of the item score patterns. We checked the assumptions of the MMH by inspecting the $H$ and $H_i$ coefficients and by inspecting estimates of the ISRFs.
Because we suspected on the basis of the literature (van den Bergh & van Ranst, 1998) and on the basis of our observations during test administration that there may be differences in the fit of the IRT model for children of different age groups and between boys and girls, we compared the fit of the model for the children age 8 and 9 ($n = 266$; hereafter called the “young children”) with children between age 10 through 12 ($n = 345$; hereafter called “old children”) and between boys and girls. By means of MSP5.0, it is possible to split the data file and investigate the $H$ values across the different groups. This was done to facilitate the interpretation of misfitting item score patterns. When items are less well scalable for particular groups of persons, individual misfit across groups may be more difficult to interpret. Furthermore, we investigated whether scale scores were similarly distributed across young and old children, and across boys and girls.

**Person Fit**

In the parametric IRT literature, there has been a proliferation of statistics and statistical tests that are suited to identify item score patterns that are improbable given an IRT model. In a nonparametric IRT context, however, there are no statistics for which a theoretical distribution is known on the basis of which an item score pattern can be classified as misfitting (Meijer & Sijtsma, 2001). Therefore, in this study, we restricted ourselves to descriptive and diagnostic information. Furthermore, we readministered the SPPC for children with aberrant item score patterns. We used a standardized version of the number of Guttman errors ($Z_{GE}$) as given by the computer program MSP5.0 as an indicator for item score pattern scalability. Guttman errors were computed for each subscale and then summed over subscales. $Z_{GE}$ is based on the ordinary transformation to $Z$ scores. $Z_{GE}$ values are usually larger for middle-scoring individuals and smaller for high- and low-scoring persons. For example, if a person receives a score of 28 out of a possible 30 points on a measure with dichotomous item scores, the number of possible Guttman errors is relatively small (between 0 and 56). On the other hand, if a person scores 15 out of 30, a large number of Guttman errors are possible (between 0 and 225). Because the

---

1 An alternative would have been to use a multidimensional IRT model and a multidimensional person-fit statistic. We did not use a multidimensional IRT model because there is not much experience with the application of multidimensional IRT models to personality data. Another reason is the relative complexity to understand the outcomes of these models for the school teachers that were involved in this project. Therefore, we used a relatively simple IRT model to get a first impression about data quality and quality of individual item score patterns. We do not expect that results would have been dramatically different had we used more complex models, but future research may use these models to answer that question.
numerical values of $Z_{GE}$ depend on the total score, we also computed the normed Guttman errors (Emons, 2008) and compared classification results on the basis of both statistics. Because there were only very small differences, we preferred to use the simplest statistic that can be computed routinely using MSP5.0.

Readmission of the SPPC

The SPPC was readministered to children with $Z_{GE}$ scores larger than 2.0. This value was based on the empirical distribution of $Z_{GE}$ scores resulting from the first administration (see Figure 3.4; upper panel) and our visual inspection of the item score patterns with $Z_{GE}$ values in the right tail of the distribution. We considered $Z_{GE}$ values smaller than 2.0 to be less “strange” than $Z_{GE}$ values larger than 2.0. This value also corresponds with the value we found when applying the rule of thumb for outliers of 1.5 times the difference between the first and the third quartile (Tukey, 1977, pp. 43-44). On the basis of regression toward the mean we expect that, in general, larger $Z_{GE}$ values would be smaller at the second administration than at the first administration. Therefore, we were especially interested in children that obtained approximately similar $Z_{GE}$ scores at both test administrations.

3.4 Results

3.4.1 Descriptive Statistics and Scalability

Scale means, standard deviations, and coefficient alpha for each of the SPPC scales are reported in Table 3.1. Except for the BC scale, on the other four scales, mean scores for the boys were higher than those for the girls. Except for the SA scale, young children scored somewhat higher than old children. Statistical $t$ tests showed that, except for the SA scale, the other mean scores differed significantly (5% level) for boys and girls. Significant age differences were only found for the SC and BC scales. Cohen’s $d$ (Cohen, 1988) values indicate that all differences were small in size.

Coefficient alpha was higher for girls than for boys and higher for old children than for young children. $H$ and $H_i$ values for the whole group are reported in Table 3.2. From the table, it can be concluded that with a few exceptions (in particular for the AC subscale), $H_i$ and $H$ values were larger than .3. Thus, most items complied with the minimum requirements of the MMH model. Further inspection of the ISRFs showed no significant violations against increasing ISRFs.
Table 3.1  
Mean, standard deviation, and coefficient alpha for each subscale of the Self-Perception Profile for Children (SPPC) for boys and girls, young and old children, and the whole sample.

<table>
<thead>
<tr>
<th>Scale</th>
<th>M</th>
<th>Cohen's d</th>
<th>Coefficient alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Young</td>
</tr>
<tr>
<td>SC</td>
<td>17.6</td>
<td>16.9</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>(3.85)</td>
<td>(3.77)</td>
<td>(3.98)</td>
</tr>
<tr>
<td>SA</td>
<td>19.0</td>
<td>18.5</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>(3.71)</td>
<td>(3.88)</td>
<td>(3.87)</td>
</tr>
<tr>
<td>AC</td>
<td>18.6</td>
<td>17.9</td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td>(3.39)</td>
<td>(3.67)</td>
<td>(3.61)</td>
</tr>
<tr>
<td>PA</td>
<td>20.4</td>
<td>19.2</td>
<td>20.2</td>
</tr>
<tr>
<td></td>
<td>(3.75)</td>
<td>(4.51)</td>
<td>(4.21)</td>
</tr>
<tr>
<td>BC</td>
<td>17.8</td>
<td>18.4</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>(3.52)</td>
<td>(3.29)</td>
<td>(3.49)</td>
</tr>
<tr>
<td>GS</td>
<td>20.7</td>
<td>20.1</td>
<td>20.5</td>
</tr>
<tr>
<td></td>
<td>(3.03)</td>
<td>(3.67)</td>
<td>(3.51)</td>
</tr>
</tbody>
</table>

*Note.* Young = young children; Old = old children; SC = Scholastic Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.
### Table 3.2

$H_i$ and $H$ coefficients of scalability for the six subscales of the SPPC for the whole sample.

<table>
<thead>
<tr>
<th>Item</th>
<th>SC</th>
<th>SA</th>
<th>AC</th>
<th>PA</th>
<th>BC</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>.35</td>
<td>.37</td>
<td>.24</td>
<td>.46</td>
<td>.24</td>
<td>.37</td>
</tr>
<tr>
<td>Item 2</td>
<td>.36</td>
<td>.45</td>
<td>.26</td>
<td>.39</td>
<td>.36</td>
<td>.36</td>
</tr>
<tr>
<td>Item 3</td>
<td>.41</td>
<td>.39</td>
<td>.26</td>
<td>.50</td>
<td>.27</td>
<td>.41</td>
</tr>
<tr>
<td>Item 4</td>
<td>.37</td>
<td>.31</td>
<td>.33</td>
<td>.53</td>
<td>.30</td>
<td>.44</td>
</tr>
<tr>
<td>Item 5</td>
<td>.43</td>
<td>.36</td>
<td>.22</td>
<td>.45</td>
<td>.37</td>
<td>.43</td>
</tr>
<tr>
<td>Item 6</td>
<td>.44</td>
<td>.34</td>
<td>.28</td>
<td>.53</td>
<td>.40</td>
<td>.32</td>
</tr>
</tbody>
</table>

$H$ coefficient | .39 | .37 | .27 | .47 | .32 | .39 |

Note. SC = Scholastic Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.

In addition, we inspected the $H$ and $H_i$ values for boys and girls, and young and old children. Scalability was better (i.e., reflected by higher $H$ values) for the old children on all scales. Scalability was better for boys on the BC scale, equal for boys and girls on the SC scale, and better for girls on the other three scales. For old children, we found for most scales medium scalability ($H$ between .40 and .50); whereas for young children, we found weak scales ($H$ between .30 and .40). For girls, except for the AC and BC scales, we found mostly medium to strong scales. For boys we found mostly weak scales.

These findings are complementary to the findings by van den Bergh and van Ranst (1998) and suggest, that in particular for young children (8 and 9 years of age), one should be careful in interpreting total scores. They suggested using separate gender and age groups when evaluating children according to their total scores. A further refinement is to evaluate the configuration of item score patterns to identify children that filled out the SPPC in an idiosyncratic way.

---

2 Group differences in scalability may be the result of differential item functioning (DIF). We investigated DIF by means of MSP5.0 for young and old children, and girls and boys. There were only small differences in the item mean scores for the different groups. Furthermore, several studies (e.g., Meijer & van Krimpen-Stoop, 2001) have shown that there was no effect of DIF on the classification of an item score pattern as fitting or misfitting; that is, person-fit results were similar.
3.4.2 Person Fit Across Different Groups

Observation

The distribution of the $Z_{GE}$ values in the full sample (Figure 3.4; upper panel) was skewed to the right. Figure 3.4 also presents the distribution for young children (lower left panel) and old children (lower right panel). As expected, young children on average have higher $Z_{GE}$ values, indicating more inconsistent behavior, than old children. We found that the means of $Z_{GE}$ for young children and old children differed significantly (means equal to .265 and -.211, respectively; $t = 5.82, p = .001$). The effect size was $\frac{1}{2}$ SD.

Figure 3.4: Histograms of the standardized number of Guttman errors ($Z_{GE}$) for the full sample (upper panel) and for young children (lower left panel) and old children (lower right panel).
Explanation

Several reasons may be suggested for this finding. Asking children to select personality statements that better describe them may be relatively complex especially for young children. They should understand the meaning of these statements, and they should also have a frame of reference that is similar to that of old children. We observed that the meaning of some items (in the Dutch translation we used) was problematic. During test administration, young children asked questions about the meaning of Item 6 of the SA scale (“some kids are popular with peers, others are not so popular”; the Dutch translation uses the word “geliefd” for “popular” [a literal translation back to English is “beloved”]; in a Dutch-language context, this word may be too difficult for young children to understand), Item 3 of the BC scale (“some kids act as is expected, others often do not act as is expected”), and Item 3 of the GS scale (“some kids tend to be self-satisfied, others often are not content with themselves”). We found large differences for young and old children between the $Hi$ values for Items 4, 5, and 6 of the SA scale and Items 1, 3, and 5 of the GS scale. This observation is important because it may partly explain why young children produce inconsistent item score patterns. When young children encounter items of which the phrasing is too difficult, they may be tempted to respond at random or they may encounter motivational problems to fill out the questionnaire.

The bipolar formulation of the questions may also be too complex for young children (see e.g., van den Bergh, 1999). When a child fills out the SPPC, he or she first has to choose between two statements. For example, “some kids often forget what they learn”, but “other kids can remember things easily”. We noticed that some children find it difficult to understand that they have to apply these statements to themselves and not to other children. Although for very young children (6 and 7 years of age), age-appropriate assessments are used in which items are formulated more concretely, and the bipolar formulation of the questions is eliminated (see e.g., van den Bergh & De Rycke, 2003), our results suggest that also for at least some children aged 8 and 9 the formulation of the questions may be too complex. Wichstrøm (1995) noted that even children and adolescents between 13 and 20 years of age sometimes only checked one side on each item.

Another explanation may be found in self-concept theory. An important question in self-concept theory and research with young children is how the dimensionality of self-concept responses for young children varies with age. Some researchers assume that self-concept is not well differentiated in early childhood (e.g., Harter & Pike, 1984; Stipek, 1981; Stipek & MacIver, 1989), whereas other researchers have found an increasing differentiation of self-concept dimensions with
age (e.g., Eccles, Wigfield, Harold, & Blumenfield, 1993; Eder, 1990; Marsh, Craven, & Debus, 1991, 1998). We hypothesize that for some young children, response behavior on a self-report inventory may be less consistent than for old children as a result of a less developed and differentiated self-concept. This resembles what Tellegen (1988) and Reise and Waller (1993) called traitedness.

Table 3.3

Proportion of boys and girls by age that fall in each response category.

<table>
<thead>
<tr>
<th>Response Category</th>
<th>Boys</th>
<th></th>
<th>Girls</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young a</td>
<td>Old b</td>
<td>Young c</td>
<td>Old d</td>
</tr>
<tr>
<td>1</td>
<td>.101</td>
<td>.064</td>
<td>.097</td>
<td>.073</td>
</tr>
<tr>
<td>2</td>
<td>.141</td>
<td>.159</td>
<td>.144</td>
<td>.201</td>
</tr>
<tr>
<td>3</td>
<td>.252</td>
<td>.332</td>
<td>.259</td>
<td>.359</td>
</tr>
<tr>
<td>4</td>
<td>.507</td>
<td>.445</td>
<td>.500</td>
<td>.367</td>
</tr>
</tbody>
</table>

a \( n = 119 \), b \( n = 140 \), c \( n = 147 \), d \( n = 194 \).

Interesting in this respect is the shift that occurs in the proportions of endorsed response categories across age and gender. Table 3.3 shows that old children more often than young children choose the Categories 2 and 3. Moreover, old girls more often than old boys choose the Categories 2 and 3. We speculate that these shifts point at a more differentiated self-concept for old children as compared to young children and at a more differentiated self-concept for girls than for boys. Jacobs, Lanza, Osgood, Eccles, and Wigfield (2002) gave three explanations why children aged 7 and 8 might exhibit extreme response behavior: unrealistically high perceptions, not being able to make use of social comparison, and limited opportunities for comparison. Furthermore, note that the frequency of misfit will be less for groups that more often choose Categories 2 and 3 than for groups that more often choose Categories 1 and 4. To further investigate unexpected answering behavior, we studied the individual item score patterns.

### 3.4.3 Person Fit at the Individual Level

Researchers typically administer personality inventories because they are interested in creating profiles of trait scores that can be used to diagnose, council, or predict behavior. Then additional information obtained from studying the configuration of item scores is useful especially when similar score profiles are the result of very different configurations of item scores. Therefore, we depicted profiles of children
with similar trait scores but different $Z_{GE}$ scores for boys (Figure 3.5; upper panel) and girls (Figure 3.5; lower panel).

Figure 3.5: Three approximately similar SPPC score profiles with different standardized number of Guttman errors ($Z_{GE}$) for boys (upper panel) and girls (lower panel). Note. $Z_{GE}$ no. 275 = 3.32, $Z_{GE}$ no. 94 = .27, $Z_{GE}$ no. 242 = -1.11; $Z_{GE}$ no. 145 = 2.09, $Z_{GE}$ no. 325 = .33, $Z_{GE}$ no. 461 = -.45. SC = Scholastic Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.

A researcher would (correctly) conclude on the basis of the three profiles in Figure 3.5 (upper panel) that all three boys have similar scores on all SPPC scales. However, it is questionable whether these similar scores have the same meaning for these three children, that is, whether these scores adequately reflect the traits being measured. Child 275 produced a very inconsistent item score pattern that consists of

We inspected the 35 item score patterns in the tail of the $Z_{GE}$ distribution with values larger than 2.0. In Table 3.4, we show the 10 most aberrant response patterns in the sample, arranged according to their $Z_{GE}$ score, in which Child 581 has the highest score. A general trend in these score patterns is that there is an unexpected alternation of 1 and 4 scores. Take Child 26; this boy produces three 4 scores and three 1 scores on the same set of items measuring social acceptance. For example, he indicates that he has as many friends as he would like to have (Item 3), and he also indicates that he would like that more children liked him (Item 5).

Table 3.4

<table>
<thead>
<tr>
<th>Child</th>
<th>Gender</th>
<th>Age</th>
<th>$Z_{GE}$</th>
<th>SPC domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SC</td>
<td>SA</td>
</tr>
<tr>
<td>581</td>
<td>boy</td>
<td>12.8</td>
<td>4.45</td>
<td>444111</td>
</tr>
<tr>
<td>26</td>
<td>boy</td>
<td>9.7</td>
<td>4.01</td>
<td>444144</td>
</tr>
<tr>
<td>12</td>
<td>girl</td>
<td>9.4</td>
<td>3.88</td>
<td>114444</td>
</tr>
<tr>
<td>411</td>
<td>boy</td>
<td>8.9</td>
<td>3.76</td>
<td>311231</td>
</tr>
<tr>
<td>306</td>
<td>girl</td>
<td>10.1</td>
<td>3.57</td>
<td>144444</td>
</tr>
<tr>
<td>538</td>
<td>boy</td>
<td>11.2</td>
<td>3.48</td>
<td>441443</td>
</tr>
<tr>
<td>137</td>
<td>boy</td>
<td>10.7</td>
<td>3.41</td>
<td>444441</td>
</tr>
<tr>
<td>275</td>
<td>boy</td>
<td>8.1</td>
<td>3.32</td>
<td>422124</td>
</tr>
<tr>
<td>101</td>
<td>boy</td>
<td>9.4</td>
<td>3.22</td>
<td>432131</td>
</tr>
<tr>
<td>515</td>
<td>girl</td>
<td>8.0</td>
<td>3.07</td>
<td>141141</td>
</tr>
</tbody>
</table>

Note. Age = age in years; $Z_{GE}$ = standardized version of the number of Guttman errors ($Z$ score); SC = Scholastic Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.
Table 3.5
Teacher-based explanations for the inconsistent answering behavior of the 35 children with initial standardized Guttman errors ($Z_{GE}$) > 2.0 (retest $Z_{GE}$ scores are also given).

<table>
<thead>
<tr>
<th>Child</th>
<th>Age</th>
<th>In.</th>
<th>Re.</th>
<th>Diff.</th>
<th>Teachers explanation the first time</th>
</tr>
</thead>
<tbody>
<tr>
<td>boy 581</td>
<td>12.80</td>
<td>4.45</td>
<td>n.a.</td>
<td>learning disability; reading comprehension; lexical processing</td>
<td></td>
</tr>
<tr>
<td>boy 26</td>
<td>9.67</td>
<td>4.01</td>
<td>-0.12</td>
<td>4.13</td>
<td>works (very) fast and inaccurate</td>
</tr>
<tr>
<td>girl 12</td>
<td>9.42</td>
<td>3.88</td>
<td>1.04</td>
<td>2.84</td>
<td>reading comprehension; lexical processing; different culture</td>
</tr>
<tr>
<td>boy 411</td>
<td>8.92</td>
<td>3.76</td>
<td>1.37</td>
<td>2.39</td>
<td>concentration; child works fast and inaccurate</td>
</tr>
<tr>
<td><strong>girl 306</strong></td>
<td>10.08</td>
<td>3.57</td>
<td>3.29</td>
<td>0.28</td>
<td>learning disability; reading comprehension; lexical processing; different culture</td>
</tr>
<tr>
<td>boy 538</td>
<td>11.17</td>
<td>3.48</td>
<td>n.a.</td>
<td>learning disability; different culture</td>
<td></td>
</tr>
<tr>
<td>boy 137</td>
<td>10.67</td>
<td>3.41</td>
<td>1.77</td>
<td>1.64</td>
<td>perhaps the combination of honest and social desirable answering; child has dyslexia</td>
</tr>
<tr>
<td>boy 275</td>
<td>8.08</td>
<td>3.32</td>
<td>1.88</td>
<td>1.44</td>
<td>reading comprehension plus child is ‘easy-going’ and has a motivation problem</td>
</tr>
<tr>
<td>boy 101</td>
<td>9.42</td>
<td>3.22</td>
<td>1.54</td>
<td>1.68</td>
<td>reading comprehension; lexical processing</td>
</tr>
<tr>
<td>girl 515</td>
<td>8.00</td>
<td>3.07</td>
<td>n.a.</td>
<td>child is very insecure and pessimistic caused by problems at home</td>
<td></td>
</tr>
<tr>
<td>boy 38</td>
<td>11.92</td>
<td>2.97</td>
<td>n.a.</td>
<td>concentration; child works fast and inaccurate</td>
<td></td>
</tr>
<tr>
<td>boy 435</td>
<td>9.33</td>
<td>2.94</td>
<td>0.67</td>
<td>2.27</td>
<td>concentration; different culture</td>
</tr>
<tr>
<td>girl 113</td>
<td>9.83</td>
<td>2.82</td>
<td>-1.27</td>
<td>4.09</td>
<td>situation at home; child is not always honest</td>
</tr>
<tr>
<td>boy 513</td>
<td>8.50</td>
<td>2.78</td>
<td>n.a.</td>
<td>reading comprehension</td>
<td></td>
</tr>
<tr>
<td><strong>boy 2</strong></td>
<td>8.08</td>
<td>2.66</td>
<td>2.55</td>
<td>0.11</td>
<td>no explanation</td>
</tr>
<tr>
<td>girl 418</td>
<td>9.00</td>
<td>2.60</td>
<td>1.26</td>
<td>1.34</td>
<td>reading comprehension</td>
</tr>
<tr>
<td>boy 532</td>
<td>9.67</td>
<td>2.56</td>
<td>n.a.</td>
<td>reading comprehension; lexical processing</td>
<td></td>
</tr>
<tr>
<td>boy 130</td>
<td>9.92</td>
<td>2.53</td>
<td>1.57</td>
<td>0.96</td>
<td>reading comprehension; concentration</td>
</tr>
<tr>
<td>girl 257</td>
<td>8.50</td>
<td>2.53</td>
<td>1.57</td>
<td>0.96</td>
<td>learning disability; reading comprehension; lexical processing</td>
</tr>
<tr>
<td>girl 393</td>
<td>7.92</td>
<td>2.47</td>
<td>0.59</td>
<td>1.88</td>
<td>reading comprehension; different culture</td>
</tr>
<tr>
<td>girl 9</td>
<td>8.25</td>
<td>2.44</td>
<td>1.29</td>
<td>1.15</td>
<td>different culture</td>
</tr>
<tr>
<td>girl 312</td>
<td>9.58</td>
<td>2.38</td>
<td>-0.62</td>
<td>3.00</td>
<td>no explanation</td>
</tr>
<tr>
<td>girl 360</td>
<td>11.17</td>
<td>2.38</td>
<td>0.30</td>
<td>2.08</td>
<td>learning disability; reading comprehension</td>
</tr>
<tr>
<td>girl 74</td>
<td>8.08</td>
<td>2.31</td>
<td>0.02</td>
<td>2.29</td>
<td>reading comprehension; lexical processing; concentration</td>
</tr>
<tr>
<td><strong>boy 144</strong></td>
<td>10.50</td>
<td>2.31</td>
<td>5.14</td>
<td>-2.83</td>
<td>reading comprehension; concentration</td>
</tr>
<tr>
<td>girl 40</td>
<td>11.00</td>
<td>2.28</td>
<td>n.a.</td>
<td>learning disability</td>
<td></td>
</tr>
<tr>
<td><strong>boy 25</strong></td>
<td>10.17</td>
<td>2.19</td>
<td>3.29</td>
<td>-1.10</td>
<td>works (very) fast and inaccurate</td>
</tr>
<tr>
<td>girl 291</td>
<td>8.83</td>
<td>2.16</td>
<td>-0.71</td>
<td>2.87</td>
<td>learning disability; reading comprehension; lexical processing</td>
</tr>
<tr>
<td>girl 409</td>
<td>9.17</td>
<td>2.16</td>
<td>0.39</td>
<td>1.77</td>
<td>reading comprehension</td>
</tr>
<tr>
<td><strong>boy 138</strong></td>
<td>10.00</td>
<td>2.13</td>
<td>3.62</td>
<td>-1.49</td>
<td>home situation</td>
</tr>
<tr>
<td>boy 45</td>
<td>11.08</td>
<td>2.09</td>
<td>1.09</td>
<td>1.00</td>
<td>learning disability; reading comprehension</td>
</tr>
<tr>
<td><strong>girl 145</strong></td>
<td>9.75</td>
<td>2.09</td>
<td>2.08</td>
<td>0.01</td>
<td>reading comprehension; concentration</td>
</tr>
<tr>
<td>boy 521</td>
<td>8.83</td>
<td>2.09</td>
<td>n.a.</td>
<td>concentration</td>
<td></td>
</tr>
<tr>
<td><strong>boy 392</strong></td>
<td>7.92</td>
<td>2.09</td>
<td>2.47</td>
<td>-0.38</td>
<td>reading comprehension; lexical processing; concentration</td>
</tr>
<tr>
<td>boy 91</td>
<td>9.42</td>
<td>2.06</td>
<td>3.17</td>
<td>-1.11</td>
<td>situation at home</td>
</tr>
</tbody>
</table>

Note. In. = $Z_{GE}$ score at the initial test administration; Re. = $Z_{GE}$ score at the retest administration; Diff. = difference between the initial and retest $Z_{GE}$ score; n.a. = not available; reading comprehension = problems with reading comprehension skills; lexical processing = problems with lexical processing speed; concentration = concentration problems. Numbers in bold identify persons who produced an inconsistent pattern at the retest administration as well as at the initial test administration.
In the Appendix, we report some observations during test taking. From these observations, we conclude that some children are less inclined to be consistent in their answering behavior and perhaps lack the cognitive ability to understand the format of the questionnaire. We further validated this conclusion by interviewing the teachers.

Table 3.5 shows the explanation given by the teachers for the inconsistent behavior of the 35 children with the most aberrant response patterns. For 21 out of the 35 children, an explanation for their inconsistent answering behavior may be lack of cognitive skills to understand the questionnaire resulting in an idiosyncratic interpretation of the SPPC scores.

### 3.4.4 Readministration of the SPPC

Thirty-five children had $Z_{GE} > 2.0$. Unfortunately, one school did not allow us to readminister the SPPC, which resulted in the loss of 8 score patterns. We readministered the SPPC to the remaining 27 children in an effort to obtain a score that may be more representative than the score obtained at the first test administration. We computed the person-fit statistics on these new data to determine consistency of answering behavior. Misfit was assessed using the item ordering obtained in the first administration. The $Z_{GE}$ values obtained from the second administration are also given in Table 3.5 (under the heading “Retest”) together with the $Z_{GE}$ values obtained at the first administration (heading “Initial”). The difference between the two $Z_{GE}$ scores (heading “Diff”) can be found in the sixth column in Table 3.5. Values close to 0 and negative values indicate that the child also produced an inconsistent pattern at the second administration. The identification numbers of these persons are given in bold in Table 3.5.

As expected, the $Z_{GE}$ scores collected at the second administration were lower than the $Z_{GE}$ scores collected at the first administration. Two observations are important here. First, 8 out of the 27 children again produced irregular item score patterns. This suggests some serious problem filling out the questionnaire. A tentative conclusion is that for 4 out of the 8 children (Children 306, 144, 145, and 392), this is due to cognitive problems: learning disability, problems with reading comprehension skills, and/or lexical processing speed. For two other children, this may be due to the home situation. Children 25 and 91 came from troubled homes, they had difficult relations with their parents, and perhaps as a result of this, they were very insecure. This may be reflected by the way they filled out the questionnaire. Harter (1983) and Leung and Leung (1992) have suggested that environmental context besides age changes influence a child’s self-concept.
Chapter 3

Child 26 presents an example of an item score pattern that was very inconsistent the first time ($Z_{GE} = 4.01$) but consistent the second time ($Z_{GE} = -0.11$). This boy worked very fast and inaccurately during the first administration. The second time the SPPC was administered in smaller groups and because of the resulting greater group pressure he filled out the SPPC more accurately. Another interesting example is Child 393. During readministration, I. J. L. Egberink found out that this girl did not understand how to answer the questions. This girl answered the questions with her classmates in mind (probably because of the phrasing: “some children like to …”, “whereas other children like to …”) instead of reporting her own functioning. So I. J. L. Egberink, in an unobtrusive way, stimulated the girl to answer the questions once more but now according to instruction. This time a consistent answer pattern resulted. Her teacher explained that this girl had weak reading comprehension skills and showed poor academic performance.

Based on the interpretation of the overall results, a general picture arises that patterns that are not in agreement with the IRT model often are the result of not understanding how to fill out the questionnaire, not understanding the meaning of statements due to low cognitive ability, and/or lack of a consistent self-concept (see Table 3.5). This conclusion is consistent with findings reported in the literature. Cramer (1999) found a linear relation between ego development and intelligence, and Marsh and MacDonald Holmes (1990) reported that children who filled out the SPPC incorrectly had a lower self-concept and showed poorer academic performance. Thus, lower level of intelligence may explain inconsistent answering behavior through (a) not understanding the format and the meaning of the items and (b) a poorly developed self-concept. As a result, a researcher should be careful in interpreting the total scores on the subscales for these children.

3.5 Discussion

In this study, we investigated the fit of an IRT model to individual score patterns on a self-report inventory and tried to interpret the results by means of additional information from observation and interviews. By means of these additional data, we obtained more insight into possible causes that underlie inconsistent response behavior. In clinical practice and applied research, the fundamental question often is not whether unexpected item score patterns exist but whether the patterns have any theoretical or applied validity. Because nothing in a person-fit procedure guarantees that identified patterns have associations with external criteria or diagnostic categories, it is important to use information from other sources. Thus, a take-home
message is that one should always combine information from person-fit statistics with information obtained from subtest scores (score profiles), interviews, and observation.

Misfitting item score patterns may also influence the fit of a test model. Recently, Woods (2006; see also Tatsuoka & Tatsuoka, 1982) proposed to use person-fit statistics to identify misfitting item score patterns and to remove them from a data set prior to conducting, for example, factor analysis. Woods showed that misfitting item score patterns may cause rejecting a one factor model for a unidimensional scale. Although we think this is an interesting application, one should be careful not to adapt the data to the test model.

Reise and Waller (1993) suggested using person-fit indices to analyze response patterns from self-report inventories. They hypothesized that untraitedness may be an explanation for misfitting response behavior. We found some evidence that inconsistent response behavior may be the result of a less-differentiated self-concept because inconsistent response behavior was more often found for young children than old children. Compared to old children, young children are not as able to evaluate the postulates of their self-concept from the standpoint of whether they are internally consistent (e.g., Harter, 1990). Furthermore, we also showed that old children suffering from learning disabilities may give extreme responses. For these children, the observed response differs from the expected response on the basis of an IRT model, and this may have important consequences for the diagnostic process.

Although person-fit statistics are not very sensitive to detect aberrant response behavior for persons with extreme (low or high) scores, in practice, this is not much of a problem because these persons are flagged by studying the subtest score profiles. For example, children with extreme scores on one or more of the individual subscales of the SPPC will fall outside the range of the subscale’s 85th and 15th percentiles. Such a profile generally would be studied in more detail by a clinician.

It is surprising that in high-stakes psychological testing, item score patterns are not screened routinely with respect to their consistency with the underlying latent trait. Large instruments such as the MMPI-2 contain response consistency scales, but many smaller scales lack such devices. Piedmont et al. (2000) remarked that with respect to the use of validity scales in personality self-report assessment, “The best evidence on protocol validity, and the best alternative to the use of validity scales, comes from the comparison of self-report scores with independent assessments, on a case-by-case basis” (p. 590). We showed that using some simple indices together with information from observation and interviews can help a researcher to decide
whether he or she can trust the score profiles on the questionnaire. Zickar and Drasgow (1996) used person-fit indices in the context of personality assessment to detect faking. In addition, we demonstrated that person-fit methods may detect other types of aberrant response behavior. In this study, some children did not understand the meaning and/or had a less differentiated self-concept and consequently, the SPPC was not suited to measure their self-concept.

Using auxiliary information from interviews and personality theory, we tried to obtain some insight into the reasons why children produce atypical response patterns. For some children, we could not explain why the patterns were unexpected. This illustrates that “there remains a gap (...) between the organized world of a mathematical measurement model and the messy world of real people reacting to a real set of items” (Molenaar, 2001, p. 295).

3.6 Some Practical Considerations for the Clinician

Finally, we discuss how person-fit measures might be used in practice. We think that routinely computing a person-fit statistic to obtain an idea about the consistency of item score answering is to be recommended3. These measures can then be combined with other information. In practice, self-report personality inventories are almost never the only source of information available to a clinician or personality researcher (and if they are, they are not very useful). Self-report personality questionnaires are often combined with cognitive measures and/or information from observation during test taking or during interviews. Data from these additional sources can be used to explain aberrant response behavior. For example, in this study, we interviewed the children’s teachers to find out whether some children were behind in their cognitive development. An alternative would have been to take the scores on a cognitive test battery or the children’s school marks. The best strategy depends on the characteristics of the specific assessment situation and also on the population of test takers. With respect to the latter, we used a questionnaire for children. In general, children aged 8 through 10 have difficulty in reflecting on their test behavior, but in our experience, older children can tell a clinician or personality researcher why they filled out an inventory as they did (as we also illustrated previously). Thus, oral reports from test takers themselves may also be an important source of information. In conclusion, combining different sources of information

3 We used the program MSP5.0. Further information about this program can be obtained from K. Sijtsma (k.sijtsma@uvt.nl) or the W. H. M. Emons (w.h.m.emons@uvt.nl). See also http://www.scienceplus.nl/scienceplus/main/softwareshop/msp.jsp
both from the self, from others, and from other questionnaires or tests may give the clinician a good picture why someone produces atypical response patterns.

3.7 Appendix: Observations During Test Administration

After instruction how to fill out the SPPC, we observed that young children, in general, filled out the SPPC faster than old children. This was surprising because the bipolar format of the questions seemed to be relatively difficult for young children to understand. However, old children more often asked questions when they did not understand how to answer questions than young children. We obtained evidence that the response format was difficult to understand for at least some young children through the remarks of several children 8 and 9 years of age who wondered what the difference was between the options “really true for me” and “sort of true for me”.

In general, old children did not choose the option “really true for me” when confronted with questions such as “I am good at sports”. Several children said that they did not choose this option because they did not want to appear arrogant. Thus, these children did not choose the extreme response options. As a result, their item score patterns may be more consistent than those of the young children.

An interesting observation was that children 11 and 12 years of age were turning back and forth the pages of the SPPC to check for earlier answers. When confronted with questions later in the questionnaire that were similar to earlier questions, they checked earlier answers because they wanted to be consistent in their answering behavior. This behavior was not observed for the young children.

Further, we observed that young children more often than old children asked for the meaning of several words. This indicates that the meaning of some items was too difficult for young children, and inconsistent answering behavior may have been the result.
Chapter 4

Detection of Aberrant Item Score Patterns in Computerized Adaptive Testing: An Empirical Example Using the CUSUM

Abstract

The scalability of individual trait scores on a computerized adaptive test (CAT) was assessed through investigating the consistency of individual item score patterns. A sample of $N = 428$ persons completed a personality CAT as part of a career development procedure. To detect inconsistent item score patterns, we used a cumulative sum (CUSUM) procedure. Combined information from the CUSUM, other personality measures, and interviews showed that similar estimated trait values may have a different interpretation. Implications for computer-based assessment are discussed.

This chapter has been published as:
4.1 Introduction

There is a large interest in the development of computer-based tests both in the clinical and the organizational domain (e.g., Hol, Vorst, & Mellenbergh, 2005; Walter et al., 2007). Advantages of computer-based testing are increased standardization, test efficiency, and combined with the Internet, it can ease the assessment procedure considerably.

An attractive application is computerized adaptive testing (CAT; Meijer & Nering, 1999), which can be constructed within the framework of item response theory (IRT; Embretson & Reise, 2000) modeling. In IRT, the person’s trait level (denoted by the Greek letter \( \theta \)) and the item characteristics are on a common metric. This property allows items to be individually tailored to a candidate’s \( \theta \) level during test administration. Another property is that once an IRT model has been fit to an item pool a person’s \( \theta \) level and standard error (\( SE \)) can be estimated using their responses to any subset of items from that pool.

The use of IRT and CAT has become popular in the ability domain (e.g., Weiss, 2004), but also in the personality domain several applications have been discussed (Hol et al., 2005; Reise & Henson, 2000). CAT may have some interesting advantages above traditional paper-and-pencil testing. Research (Hol et al., 2005; Reise & Henson, 2000) showed substantial item savings when using CAT, while maintaining a high correlation between \( \theta \) estimates (denoted by \( \hat{\theta} \)) based on CAT and full scale \( \hat{\theta} \) s. Although some authors (e.g., Reise & Henson, 2000) concluded that test shortening (choosing the best discriminating items) may be equally efficient to CAT administration, Hol, Vorst, and Mellenbergh (2007) showed that CAT outperformed short forms, especially at the extremes of the latent trait scale.

A serious threat to the validity of \( \hat{\theta} \), however, is inconsistent response behavior. This type of behavior may result in aberrant or misfitting item response patterns which may seriously affect the interpretability of \( \hat{\theta} \) (e.g., Meijer & Nering, 1997). Meijer, Egberink, Emons, and Sijtsma (2008) analyzing an often-used inventory on self-perception showed that some children produced inconsistent response patterns as a result of problems understanding the questions. Furthermore, \( \hat{\theta} \) s may be invalid due to faking, unmotivated response behavior, or traitedness (i.e., persons that do not fit a particular conception of a latent trait, Tellegen, 1988). Inconsistent response behavior may affect the ordering of persons according to their \( \hat{\theta} \) level and, consequently, it may affect individual classification (e.g., Meijer & Sijtsma, 2001). Moreover, in a CAT inconsistent response behavior may also have detrimental effects on its efficiency because it may influence the administration order of the
items for an individual candidate. But, perhaps most important, inconsistent response behavior may interfere with the adequate communication between a psychologist and a test-taker.

For fixed tests there is a body of literature that discusses statistical methods (so-called person-fit statistics) to detect invalid item score patterns (e.g., Meijer & Sijtsma, 2001). The \( l_x \) statistic (Drasgow, Levine, & Williams, 1985) and the number of Guttman errors (e.g., Meijer, 1994) are examples of person-fit statistics. According to the Guttman model it is not allowed that a subject endorses a less popular item while rejecting a more popular item.

In the CAT literature there are only a few studies that discuss statistical methods to detect inconsistent response behavior (e.g., Armstrong & Shi, 2009; Bradlow, Weiss, & Cho, 1998; van Krimpen-Stoop & Meijer, 2002) and these studies mainly used simulated data. Thus, there is not much experience how these methods will perform using real empirical data. There are no studies that investigate why persons produce inconsistent response patterns on a CAT. Therefore, in the present study, we extend the CAT literature by investigating the validity of individual trait scores on a personality questionnaire through studying individual item score patterns using empirical data. Furthermore, we tried to get a better picture of persons who produce inconsistent response behavior through combining information obtained from statistical indices, other measurement instruments, and written interviews.

This study is organized as follows. First, we describe the construction of the personality CAT. Second, we apply statistical methods to investigate inconsistent response behavior. Finally, we show that similar trait scores may be the result of very different response behavior and we show how this information can help test score interpretation.

**4.2 Method**

**4.2.1 Development of the CAT**

**Item Pool Development**

We collected data from 984 Dutch candidates who were administered the Workplace Big Five personality questionnaire (WB5; Schakel, Smid, & Jaganjac, 2007) as part of a career development procedure. The WB5 is an online, computer-based personality questionnaire applied to situations and behavior in the workplace. It consists of 144 items, distributed over five scales (Emotional Stability, Extraversion, Openness,
Agreeableness, and Conscientiousness). The items are scored on a 5-point Likert scale. The answer most indicative of the trait is scored “5” and the answer least indicative is scored “1”. Coefficient alpha varies from .79 to .91 for the five scales. Items from the Big Five factors Conscientiousness (C), Emotional Stability (ES), and Extraversion (E) were selected for the CAT item pool. We developed these three separate CATs in collaboration with a Dutch assessment company. Based on their own assessment experience, they were most interested in measuring these three factors. From these original three scales, items were selected that allowed unidimensional measurement and discriminated well between persons.

To select items that together formed a unidimensional scale, we checked the assumptions of the Mokken model of monotone homogeneity (MMH, e.g., Sijtsma & Molenaar, 2002) using the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0, Molenaar & Sijtsma, 2000) by inspecting the $H$ coefficients for scales and the $H_i$ coefficients for items. Increasing values of $H$ and $H_i$ between .30 and 1.00 (maximum) reflect better scalability of sets of items and individual items, respectively (for a discussion of these measures see for example, Sijtsma & Molenaar, 2002). We used $H \geq .30$ as a lower bound which corresponds to lower bounds used in the literature (e.g., Meijer, Sijtsma, & Smid, 1990). The MMH can be interpreted as a nonparametric version of the graded response model (GRM; Samejima, 1969, 1997) that we used in the CAT algorithm (to be discussed below). The GRM uses a logistic function, whereas in the MMH the form of the item response function is allowed to take any form as long as the function is nondecreasing. In the literature (e.g., Embr etson & Reise, 2000; Sijtsma & Molenaar, 2002) it has been discussed that the GRM is best suited for tests using ordered response categories, such as Likert-type scales. Therefore, we used the GRM. Because the MMH is the most general model, we used it to select items.

This resulted in an item pool of 31 Emotional Stability items ($H = .31$), 27 Extraversion items ($H = .34$), and 23 Conscientiousness items ($H = .33$). Because $H \geq .30$, the scales and their items complied with the minimum requirements of the MMH model.

**CAT Selection Algorithm**

Item parameters for the GRM were estimated using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003) with marginal maximum likelihood estimation. The GRM is a 2-parameter-logistic (2PL) model for polytomous data. The model assumes that a person makes a global evaluation before responding to an item. For example, for an item with four categories, the person compares the first category with the second,
third, and fourth category; the first and second with the third and fourth category; and the first, second and third category with the fourth category. The person immediately seeks his or her position on the scale. Each item $i$ is described by a discrimination parameter ($\alpha$ parameter) and two or more location or “difficulty” parameters ($\beta_m$ parameters); the number of location parameters per item is equal to the number of response categories minus 1. The location parameter $\beta_m$ can be interpreted as the point at the latent trait continuum where there is a 50% chance of scoring in category $m$ or higher. The $\alpha$ and $\beta_m$ parameters are used to determine the probability of an examinee to respond in a particular response category (Embretson & Reise, 2000).

Important tools in the context of CAT are the item and test information curves (Embretson & Reise, 2000). The item information curve indicates the amount of psychometric information an item provides at each $\theta$ level, based on the $\alpha$ parameter and the probabilities of responding in a certain category. These curves are additive across items on a common scale and together constitute the test information, denoted by $TI(\theta)$, which indicates the amount of information a test provides at each $\theta$ level. It is inversely related to standard error of measurement. As a stopping rule we used $SE < .32$, this corresponds to a reliability $\geq .9$ for each individual.

4.2.2 Participants and Procedure

CAT data were collected as part of a career development procedure of a Dutch professional care company. There were 428 participants with a mean age of 32.1 ($SD = 12.7$); 29% mostly White men and 71% mostly White women. 28.7% of the participants had a university degree, 59.1% had higher education, and 12.1% secondary education.

Participants received an email with instructions and a hyperlink to the CAT. Eight days later they received an email with a hyperlink to the WB5. Consequently, it was possible to compare the score on the CAT scales with a full scale score based on the WB5 responses.
4.2.3 Analysis

Inconsistent Response Patterns

To investigate whether persons were inconsistent in their response behavior, we
used a method that was proposed by van Krimpen-Stoop and Meijer (2002). The
idea is that large differences between observed and expected scores based on an IRT
model may point at inconsistency. To detect inconsistent response behavior in
polytomous CAT, van Krimpen-Stoop and Meijer proposed the cumulative sum
procedure (CUSUM; for recent developments see Armstrong & Shi, 2009). In their
procedure, where \( i_k \) denotes the \( k \)th item in the CAT, with \( j = 0 \ldots m \) response
categories, statistic \( T_k \) is the residual between the observed and expected score of the
\( k \)th administered item, corrected for the number of items in the test (\( N \)):

\[
T_k = \frac{1}{N} \left[ x_k - \sum_{j=0}^{m} j P_j (\hat{\theta}_N) \right]
\]

with \( P_j (\hat{\theta}_N) \) equal to the probability of responding in category \( j \) on item \( i \),
conditional on \( \hat{\theta} \) after the CAT has been completed.

In a CUSUM, these residuals are summed across consecutive items. This is done
in a way that enables the researcher to separate strings of negative and positive
residuals. For each candidate, after each administered item \( i_k \) of the CAT, the
CUSUM is defined as follows:

\[
C_k^+ = \max \left[ 0, T_k + C_{k-1}^+ \right]
\]

\[
C_k^- = \min \left[ 0, T_k + C_{k-1}^- \right]
\]

\[
C_0^+ = C_0^- = 0
\]

where \( C^+ \) and \( C^- \) are sensitive to series of positive and negative values of \( T_k \),
respectively: \( C^+ \) becomes large only for series of consecutive positive values of \( T \),
whereas \( C^- \) becomes small only for series of consecutive negative values of \( T \).

The CUSUM procedure consists of three steps; (1) \( \hat{\theta}_N \) is estimated, (2) \( T \) is
determined for each administered item, and (3) \( C^+ \) and \( C^- \) are calculated for each
administered item. To illustrate this, consider a person taking a 20-item dichotomous
CAT who generates the response pattern given in Table 4.1. \( \hat{\theta}_N \) for this candidate
was -.221. Table 4.1 gives the values \( T, C^+, \) and \( C^- \) after the administration of each
item. To illustrate the computation of \( C^+ \) and \( C^- \), consider the first three items.
The first item score equals 0 and \( P_i (\hat{\theta}) = .411 \), which results in \( T_1 = -.021, C_1^+ = 0, \)
and $C_{1}^{-} = -.021$. The candidate also did not endorse the second item, resulting in $T_{2} = -.022$, $C_{2}^{+} = 0$, and $C_{2}^{-} = -.021 + -.022 = -.043$. The third item is endorsed, which results in $T_{3} = .025$, $C_{3}^{+} = 0 + .025 = .025$, and $C_{3}^{-} = -.042 + .025 = -.017$. Note that the procedure is running on both sides and that a negative (or positive) value contributes to $C^{+}$ and $C^{-}$, respectively. However, because the smallest value of $C^{+}$ and the largest value of $C^{-}$ equals zero, strings of positive and negative residuals can be distinguished.

Table 4.1  
CUSUM procedure for a fitting response pattern.

<table>
<thead>
<tr>
<th>Item</th>
<th>$x$</th>
<th>$P$</th>
<th>$T$</th>
<th>$C^{+}$</th>
<th>$C^{-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>.411</td>
<td>-.021</td>
<td>0</td>
<td>-.021</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>.439</td>
<td>-.022</td>
<td>0</td>
<td>-.043</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>.497</td>
<td>.025</td>
<td>.025</td>
<td>-.017</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>.476</td>
<td>-.024</td>
<td>.001</td>
<td>-.041</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.580</td>
<td>.021</td>
<td>.022</td>
<td>-.020</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>.463</td>
<td>-.023</td>
<td>0</td>
<td>-.043</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>.514</td>
<td>.024</td>
<td>.024</td>
<td>-.019</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>.578</td>
<td>-.029</td>
<td>0</td>
<td>-.048</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>.664</td>
<td>.017</td>
<td>.017</td>
<td>-.031</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>.568</td>
<td>.022</td>
<td>.038</td>
<td>-.009</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>.534</td>
<td>.023</td>
<td>.062</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>.287</td>
<td>-.014</td>
<td>.047</td>
<td>-.014</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>.424</td>
<td>-.021</td>
<td>.026</td>
<td>-.036</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>.557</td>
<td>.022</td>
<td>.048</td>
<td>-.013</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>.411</td>
<td>-.021</td>
<td>.028</td>
<td>-.034</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>.421</td>
<td>-.021</td>
<td>.007</td>
<td>-.055</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>.679</td>
<td>.016</td>
<td>.023</td>
<td>-.039</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>.418</td>
<td>.029</td>
<td>.052</td>
<td>-.010</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>.319</td>
<td>-.016</td>
<td>.036</td>
<td>-.026</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>.606</td>
<td>.020</td>
<td>.056</td>
<td>-.006</td>
</tr>
</tbody>
</table>

Note. $x =$ item score; $P =$ probability of endorsing an item given $\hat{\theta}$; $T =$ difference between observed and expected score; $C^{+}$ = maximum value of the CUSUM; $C^{-}$ = minimum value of the CUSUM.

To classify a response pattern as misfitting an appropriate upper bound (UB) and lower bound (LB) need to be set. Then, when $C^{+} > UB$ or $C^{-} < LB$ the response pattern is classified as misfitting; when $C^{+} \leq UB$ and $C^{-} \geq LB$ the response pattern is classified as fitting. Several alternatives to determine UB and LB in a CUSUM are described in van Krimpen-Stoop and Meijer (2002). In the present study, the UB was determined as the value of maximum $C^{+}$ for which 5% of the respondents had higher maximum $C^{+}$ values and the LB was determined as the value of minimum
C⁻ for which 5% of the respondents had lower minimum C⁻ values. UB values for the three scales were, 1.06 (ES), 1.02 (E), and 1.05 (C), and LB values were -.09 (ES), -.06 (E), and -.10 (C).

4.3 Results
Before we discuss the results with respect to the consistency of individual response behavior, we first discuss some characteristics of the CAT.

4.3.1 CAT and Full Scale Comparison
When comparing CAT and full scale administration, we found that the overall item saving equaled 49%. The Emotional Stability scale has the largest item savings with 51%.

When we consider the mean psychometric information per item (i.e., the psychometric efficiency), the CAT has a higher efficiency per administered item than the full scale. Reduction in administration time for the CAT as compared to the full scale was on average 52.8%, which means, on average, a reduction of more than five minutes. Emotional Stability, Extraversion, and Conscientiousness CAT scores correlated .83, .88, and .84, with the full scale scores, respectively. CAT scores correlated .83, .81, and .83 with their WB5 counterparts for Emotional Stability, Extraversion, and Conscientiousness, respectively. Thus, CAT and full scale scores, and CAT and original WB5 scale scores are highly correlated. These findings are important to interpret individual scores.

4.3.2 Detecting Person Misfit
There were 14 (ES), 14 (C), and 10 (E) persons with C⁺ > UB and there were 8 (ES), 7 (C), and 12 (E) persons with C⁻ < LB. In Figure 4.1 two empirical CUSUMs are shown belonging to the inconsistent response pattern of Person 942 (\(\hat{\theta}_N = -1.56\)) and the consistent response pattern of Person 840 (\(\hat{\theta}_N = -1.36\)). A psychologist would (correctly) conclude on the basis of their \(\hat{\theta}\) s that these persons score equally low on the Emotional Stability scale. However, it is questionable whether these similar scores have the same meaning for these persons; that is, whether these \(\hat{\theta}\) s adequately reflect the trait being measured. The response pattern of Person 942 consists of only extreme scores (511151115511155), whereas Person 840 produces a more consistent response pattern given \(\hat{\theta}_N\) (22222232222222).
Person 942 often chooses response categories that are larger than the expected response categories which results in high $C^+$ scores. Figure 4.1 shows that for the response pattern of Person 942 all $C^-$ values are larger than -.09, whereas at the 14th item the value of $C^+$ becomes larger than 1.06. As a result, this response pattern is classified as misfitting. For the response pattern of Person 840 all $C^+$ values are smaller than 1.06 and, also, all $C^-$ are smaller than -.09. Therefore, this pattern is classified as fitting.

![CUSUM patterns for Person 942 ($\hat{\theta}_N = -1.56$) and Person 840 ($\hat{\theta}_N = -1.36$).](image)

However, the fundamental question here is not whether unexpected item response patterns exist but whether the patterns have any theoretical or applied validity. Because nothing in a CUSUM procedure guarantees that identified patterns have associations with external criteria, it is important to use information from other sources such as interviews and observation.

### 4.3.3 What Explains Inconsistent Response Behavior?

Several reasons may be suggested for the inconsistent response patterns discussed above. CUSUM procedures are sensitive to response behavior that is characterized by runs of alternating extreme scores. In fact, these persons behave as if the items are dichotomously phrased. This may suggest a trait manifestation that is different from a majority of the persons in the sample. However, the problem with the interpretation of inconsistent response behavior is that it can be caused by many
factors besides lack of traitedness (Reise & Waller, 1993; Tellegen, 1988). For example, response errors and cheating limit the interpretability of CUSUM scores. Therefore, to explain the differences between normal and inconsistent response behavior, we conducted additional analyses.

First, we examined whether scalability is trait-specific by computing the correlations between the CUSUM scores across the three scales. We observed that most correlations were positive but small. Most correlations fluctuated around $r = .08$ (95% CI = [-.02, .18]). Thus, most scalability variation is trait specific. Furthermore, to examine this association more carefully, the number of persons who were consistently classified as misfitting or fitting across the three CATs was assessed (see also Reise & Waller, 1993). There were only 3 persons that were classified as inconsistent on two scales based on their $C^+$ values and 2 persons that were classified as inconsistent on two scales based on their $C^-$ values. No persons were classified as aberrant on all three scales.

Second, we correlated the trait level scores on the WB5 with the CUSUM scores on the three CATs (see Table 4.2). Interesting was that on two of the three subscales CUSUM scores correlated negatively with Agreeableness, indicating that the lower a person scored on Agreeableness, the more model-inconsistent item responses were produced. A more detailed analysis showed that especially the relation between the subscale Tact and the CUSUM scores was significant for two scales.

Table 4.2

<table>
<thead>
<tr>
<th></th>
<th>Maximum $C^+$ values</th>
<th>Minimum $C^-$ values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>E</td>
</tr>
<tr>
<td>NEU</td>
<td>.06</td>
<td>-.01</td>
</tr>
<tr>
<td>EXT</td>
<td>-.12</td>
<td>.02</td>
</tr>
<tr>
<td>OPE</td>
<td>.02</td>
<td>.06</td>
</tr>
<tr>
<td>AGR</td>
<td>-.12</td>
<td>.04</td>
</tr>
<tr>
<td>CON</td>
<td>-.18</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note. Values enclosed in brackets represent the 95% confidence intervals. NEU = Neuroticism; EXT = Extraversion; OPE = Openness; AGR = Agreeableness; CON = Conscientiousness; ES = Emotional Stability; E = Extraversion; C = Conscientiousness.
Third, we investigated whether gender was related to inconsistent response behavior. Schmitt, Chan, Sacco, McFarland, and Jennings (1999) found that females were more consistent than males. Their study, however, was conducted in the context of fixed paper-and-pencil tests using statistics that were sensitive to different kinds of aberrant behavior. In the present study, we found that both $C^+$ and $C^-$ values correlated positively with gender on the three CATs, indicating that females were more consistent than males.

Table 4.3

<table>
<thead>
<tr>
<th>Person</th>
<th>Response behavior</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>942</td>
<td>Only '1' and '5'</td>
<td>“If there are multiple response categories, I force myself to choose between the two extreme categories. For me the categories in the middle suggest that you doubt about your answer and I have no doubts. I feel choosing between the two extreme categories is most honest. Besides this, I hold uncompromising views and I do not change my opinion or view.”</td>
</tr>
<tr>
<td>807</td>
<td>Only '1' and '5'</td>
<td>“I hold strong views about certain things and stick to my views. I am also very direct to other people, this often results in conflicts. I am not surprised about my response behavior, in my opinion it matches with my personality.”</td>
</tr>
<tr>
<td>947</td>
<td>Mostly '1' and '5', and sometimes '3' or '4'</td>
<td>“I hold strong views about certain things; it is either yes or no. I always respond in an extreme manner and stick to my view. When I thought the statement was ambiguous or when the statement was not applicable to me, I choose the option in the middle, for me the ‘neutral’ option.”</td>
</tr>
</tbody>
</table>

Comment authors. She also wondered whether her husband, also a candidate, needed to give an explanation. The first author explained to her that he did not respond in such an extreme manner. She replied “He is calm and keeps an open mind. He is not extreme or very direct; he is more careful and balanced than I am.”
Fourth, we conducted a small study to obtain qualitative data through interviewing persons with inconsistent response patterns. In the person-fit literature there are almost no direct measures that may illuminate the reason for inconsistency. So, there is a need for more clinically oriented studies that follow up with respondents to investigate how they differ from typical respondents (Meijer et al., 2008; Rudner, Bracey, & Skaggs, 1996). Therefore, the first author sent emails to the 21 persons with the most extreme $C^+$ or $C^-$ values on the Conscientiousness scale in which she explained that their response pattern consisted of many extreme responses and whether they had an explanation for it. Furthermore, it was emphasized that these questions were treated confidentially and that they were only used for research purposes. We obtained written answers from 13 persons. In Table 4.3 we depicted some of their reactions. A tentative conclusion is that these candidates hold strong views and opinions about themselves and others. Thus, this response behavior seems to be related to trait specific characteristics.

4.4 Discussion

The aim of the present study was the application of the relatively new CUSUM method to detect inconsistent response patterns, which resulted in additional information to $\hat{\theta}$. We combined information obtained from statistical measures with self-report scores and interviews. A tentative conclusion on the basis of the correlations with the WB5 and the interviews is that persons with high CUSUM values produce inconsistent response patterns as a result of a tendency to seeing the world in terms of extremes. This may be very interesting additional information in an assessment procedure. Reise and Waller (1993) found that inconsistent behavior was related to negative affectivity (distrust, fearfulness, hostility). However, they used a different person-fit statistic, which may result in detecting a different type of inconsistent response behavior.

CUSUM procedures are particularly sensitive to runs of large differences between observed and expected scores given $\hat{\theta}$. Important to understand, however, is that although faked protocols will probably result in runs of extreme scores in the keyed direction, CUSUM procedures will not be very sensitive to these runs because they will also result in a mean shift of the item responses and thus in a higher $\hat{\theta}$.

In this study $LB$ and $UB$ were based on the sample distribution. An alternative is to conduct a simulation study based on the estimated item parameters from the sample. Theoretical distributions for the CUSUM we used in this study are unknown to the authors.
An interesting application for future research is the use of a CUSUM procedure during test administration. For example, when it is clear that a person is very inconsistent in his or her response behavior, items with a more subtle content may be selected. Doing so, CUSUM information will help to make the test more interactive.

At the moment we are incorporating the CUSUM procedure in an assessment procedure. Besides a total score on the CAT, a psychologist then obtains information about the consistency of the response behavior and thus about the validity of the total score. A psychologist can use this information in, for example, an interview to check why the candidate responded inconsistently. The examples sketched above illustrate that this may result in a more refined picture of a candidates response behavior.

It is clear from our study that one should be very careful in labeling a response pattern as misfitting only based on a person-fit statistic. Additional information is needed. A possible strategy may be to ask a respondent for the reasons for misfit. For example, in intelligence testing, a psychologist may ask whether a respondent was extremely nervous at the start of the test. In personality assessment, it may be possible to check by means of directed questions, as we did, whether the personality trait being measured was applicable to the respondent.
Chapter 5

Conscientiousness in the Workplace: Applying Mixture IRT to Investigate Scalability and Predictive Validity

Abstract

Mixture item response theory (IRT) models have been used to assess multidimensionality of the construct being measured and to detect different response styles for different groups. In this study a mixture version of the graded response model was applied to investigate scalability and predictive validity for a Conscientiousness scale in a career development context (N = 9283). A four-class solution yielded the most interpretable results. The classes differed mainly with respect to their scores on the subscales Perfectionism and Concentration. Results showed that Conscientiousness may be qualitatively different for different groups of persons and that the predictive validity of the test scores improved for persons in different classes as compared to fitting a unidimensional IRT model. Implications of this study for personality assessment are discussed.

This chapter has been published as:
5.1 Introduction

Conscientiousness is one of the most important personality traits in the workplace assessed by applied researchers and psychologists because it predicts different types of job performance and various outcomes related to social functioning (e.g., Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Dudley, Orvis, Lebiecki, & Cortina, 2006; Roberts, Chernyshenko, Stark, & Goldberg, 2005). Results of several studies show, however, that it is important to pay attention to the lower-order structure of Conscientiousness, in particular to assess predictor-criterion relationships in industrial and organizational (I/O) psychology. For example, Dudley et al. (2006) specified four narrow traits (achievement, order, cautiousness, and dependability) and examined the predictive power of these narrow traits. They found that narrow traits do have incremental validity above and beyond global Conscientiousness, although the incremental validity depended on the particular performance criteria and the occupation in question.

An interesting alternative to obtain more detailed information about the lower-order constructs of Conscientiousness is to consider Conscientiousness scales and subscales from an individual perspective. This would lead to a more person-centered approach instead of a variable-centered approach, which is more common in personality research. In a variable-centered approach the relation between items and subsets of items are the main focus of interest, whereas in a person-centered approach the main focus is on the person and differences between persons (e.g., differences in the use of the response scale at the individual level). From a psychological assessment point of view, it is most interesting to investigate whether there are individual differences in the way subsets of items are interpreted. This is important in personality assessment because if there are differences, this may affect individual classification and the predictive validity of individual test scores (e.g., Austin, Deary, & Egan, 2006; Weekers & Meijer, 2008).

With the increasing popularity of item response theory (IRT, e.g., Embretson & Reise, 2000) models, techniques have been proposed to analyze individual response patterns and to uncover subgroups that have different probabilities to endorse an item. One such technique is mixture IRT (e.g., Rost, 1990; Rost & Langeheine, 1997). Mixture IRT models combine latent class models and IRT models by identifying groups of individuals (i.e., latent classes, instead of manifest classes, such as gender or age) within a given sample for whom a specific IRT model is applicable. These classes differ from each other with regard to their use of the response scale. Mixture IRT models have been applied in the personality domain for different
purposes. Reise and Gomel (1995) suggested that mixture IRT models can be of help to uncover groups of persons who are qualitatively distinct with respect to the probability of endorsing a particular set of items. That is, groups of persons may differ in the way a psychological construct is applicable. Eid and Zickar (2007) used mixture IRT models in the personality domain to uncover faking and response styles. Mixture IRT models have also been applied in other research domains. For example, Eid and Rauber (2000) used these models to detect measurement invariance in organizational surveys and Muthén and Asparouhov (2006) applied these models to investigate tobacco dependence criteria.

The aim of the present study was threefold: (1) to investigate the number of classes that differ systematically in their response scale usage of a Conscientiousness scale consisting of four lower-order constructs (Perfectionism, Organization, Concentration, and Methodicalness), (2) to explore the characteristics that distinguish the particular classes, and (3) to evaluate the predictive validity of the mixture IRT trait estimates as compared to the estimates under a unidimensional IRT model by relating the estimates to an external criterion measure. Therefore, we first discuss (a) research concerning the structure of Conscientiousness and its predictive validity, (b) the principles of IRT and mixture IRT, and (c) some recent applications of mixture IRT. Second, we apply mixture IRT to a Conscientiousness scale and discuss the differences between the classes with respect to the scalability and the predictive validity of the mixture IRT trait estimates compared to the unidimensional IRT trait estimates. Finally, we reflect on the usefulness of this method for personality assessment.

5.2 The Structure and the Predictive Validity of Conscientiousness

5.2.1 Structure of Conscientiousness

There are different personality questionnaires that take Conscientiousness into account, sometimes consisting of different subscales. Also, the theoretical and empirical underpinnings of these different questionnaires may differ. Therefore, several researchers investigated the lower-order structure of Conscientiousness by analyzing item content and factor analyzing different questionnaires. For example, Saucier and Ostendorf (1999) examined the structure of 500 adjectives of the Big Five. For the Conscientiousness factor, they found four subcomponents
(orderliness, decisiveness-consistency, reliability, and industriousness). Peabody and De Raad (2002) combined the results of six studies in different languages to develop a lower-order structure for each Big Five factor. For Conscientiousness, they identified the subcomponents orderliness, work, responsibility, and impulse control as only related to Conscientiousness.

In general, different studies that examined the lower-order structure of Conscientiousness (e.g., MacCann, Duckworth, & Roberts, 2009; Peabody & De Raad, 2002; Perugini & Gallucci, 1997; Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004; Roberts et al., 2005; Saucier & Ostendorf, 1999) identified three common subcomponents, related only to Conscientiousness and not to another Big Five factor: orderliness (i.e., being neat and organized), industriousness (i.e., hard working and being ambitious), and impulse control (i.e., being careful, patient, and cautious). A sometimes identified fourth subcomponent is responsibility, also labeled as reliability and dependability (i.e., being trustworthy, responsible, and dependable). However, this subcomponent is often suggested to be a mix of Conscientiousness and Agreeableness (MacCann et al., 2009) or a mix of Conscientiousness and Emotional Stability (Roberts et al., 2005).

In the present study, we used a subscale measuring Perfectionism which is sometimes also considered a mix of Conscientiousness and Neuroticism (e.g., Roberts et al., 2005). Hamachek (1978) distinguished two forms of Perfectionism, a positive form called ‘normal perfectionism’ (persons who enjoy pursuing their perfectionistic strivings) and a negative form called ‘neurotic perfectionism’ (persons who suffer from their perfectionistic strivings).

5.2.2 Predictive Validity of Conscientiousness

In I/O psychology several researchers noticed that it is important to pay attention to the lower-order structure of the construct of interest when assessing predictor-criterion relationships (see Hough & Ones, 2001; Hough & Oswald, 2000; Paunonen & Ashton, 2001; Paunonen, Haddock, Forsterling, & Keinonen, 2003). In a meta-analysis, Hurtz and Donovan (2000) found an average corrected criterion-related validity between global Conscientiousness and job performance of $r = .22$. Although broad trait measures generally maximize prediction of overall job performance, narrow trait measures maximize the predictive validity of specific criteria. In order to maximize the validity of narrow traits, traits must be selected on the basis of strong a priori linkages to a criterion. Furthermore, narrow traits may help to understand the personality-based causes of individual differences in working behavior (e.g., Hough & Ones, 2001).
Roberts et al. (2005) compared the criterion-related validity of the six subscales they found (industriousness, order, self-control, responsibility, traditionalism, and virtue) to the criterion-related validity of the overall Conscientiousness scale. Results showed that the subscales had a higher criterion-related validity than the Conscientiousness scale in nearly all cases (see also Dudley et al., 2006).

Recently, MacCann et al. (2009) examined the lower-order structure of 18 IPIP scales relating to Conscientiousness (i.e., 117 items). Their analyses resulted in eight subscales: industriousness, perfectionism, tidiness, procrastination refrainment, control, cautiousness, task planning and perseverance. Only control and perseverance were also related to another Big Five factor, namely Agreeableness and Neuroticism, respectively. When examining the criterion-related validity of the broad Conscientiousness factor compared to the validity of the eight different subscales, only perfectionism and industriousness had a significantly higher relationship to the criteria than Conscientiousness. Perfectionism yielded a stronger relationship with Secondary School Admission Test percentiles and industriousness with absence from class.

5.3 Item Response Theory and Mixture Item Response Theory

5.3.1 Item Response Theory

Item response theory (e.g., Embretson & Reise, 2000) is a collection of statistical models that can be used to analyze items and scales, to create and administer psychological measures, and to measure individuals on psychological constructs. In most IRT models, test responses are assumed to be influenced by a single latent trait, denoted by the Greek letter θ. For dichotomous (true, false) data, the goal of fitting an IRT model is to identify an item response function (IRF) that describes the relation between θ and the probability of item endorsement. In most IRT models, it is assumed that the probability of item endorsement should increase as the trait level increases; thus, IRFs are monotonically increasing functions.

Compared to classical test theory (CTT), IRT has a number of advantages. One advantage is that to judge the quality of an item, one can transform the item’s IRF into an item information function, which shows how much psychometric information (a number that represents an item’s ability to differentiate among persons) the item provides at each trait level. Different items can provide different
amounts of information in different ranges of a given latent trait. Item and scale information are analogous to CTT’s item and test reliability. An important difference, however, is that under an IRT framework, information (precision) can vary depending on where an individual falls along the trait range, whereas in CTT, the scale reliability (precision) is assumed to be the same for all individuals, regardless of their raw-score levels. Another advantage of IRT is that, because it is a model-based approach, it is possible to predict a person’s answering behavior when confronted with a particular set of questionnaire items. The IRF gives the probability of endorsing an item for each latent trait value.

Several authors have introduced and discussed the advantages of applying IRT models (e.g., Embretson & Reise, 2000) to construct personality scales and to explore the structure of personality data sets. Reise and Waller (2009) discussed that using classical test theory instead of IRT to analyze personality data may hide important subgroup differences. Waller, Tellegen, McDonald, and Lykken (1996) contrasted the use of IRT with principal component factor analysis and showed how traditional factor analysis can produce misleading results when applied to personality items in the sense that a linear factor analysis can produce spurious factors when applied to items that have nonlinear regressions on the underlying content factors. Therefore, we used an IRT approach in this study.

5.3.2 Mixture Item Response Theory

Since the start of psychological testing, it has been recognized that individual persons or subgroups of persons may behave differently than the majority of persons in the population. For example, early attempts to study profiles of item and test scores have been reported by Du Mas (1946), Osgood and Suci (1952), and Cronbach and Gleser (1953). Furthermore, techniques like cluster analysis (such as K-means cluster analysis, see e.g., Steinley, 2003) and discriminant function analysis have been used to identify distinct homogeneous groups. Also, on the individual level, the use of response scales like the Variable Response Inconsistency Scale and the True Response Inconsistency Scale of the Minnesota Multiphasic Personality Inventory-2 (MMPI–2; Butcher et al., 2001) or the use of person-fit indices (e.g., Meijer, Egberink, Emons, & Sijtsma, 2008) have been discussed. Person-fit indices can be used to assess how well an individual’s responses conform to the measurement model used to interpret individual differences in the trait level (e.g., Meijer & Sijtsma, 2001). An alternative technique to provide important insights into the nature of response behavior on tests is mixture IRT, which combines latent class analysis (LCA) with traditional IRT.
LCA is a multivariate technique that attempts to identify distinct classes of individuals on a psychometric scale (see Lazarsfeld & Henry, 1968). Individuals within a single class are assumed to behave similarly on relevant behavior. Members of different classes, though, are assumed to behave differently. For example, Nestadt et al. (2009) recently showed that obsessive-compulsive disorder can be classified into three classes based on co-morbidity.

Typical IRT models assume that data come from one population in which all members of that population respond to the items using the same answering process; one IRF is assumed to characterize all respondents within that sample. With mixture IRT, it is possible to identify distinct subpopulations within a larger population, each of which responds differently to a set of items (see Rost, 1990; Rost & Langeheine, 1997). Therefore, mixture IRT combines LCA with IRT by identifying groups of individuals (i.e., latent classes, instead of manifest classes, such as gender or age) within a given sample for whom a specific IRT model is applicable. Often, this means that for each latent class the same IRT model holds, but with different parameters across the latent classes. Individuals within these latent classes share the way of responding to the response scale. Based on a person’s response pattern, the conditional probability of belonging to each of the latent classes and the corresponding latent trait score in each latent class is estimated and persons are assigned to the latent class with the highest conditional probability. Mixture IRT is similar to differential item functioning (DIF) analysis (see Cohen & Bolt, 2005). However, DIF analysis often uses a priori grouping information by means of manifest variables (e.g., gender, age, or education), whereas in mixture IRT no a priori grouping information has to be used. Instead, two or more latent classes are identified that differ systematically in response scale usage. Sometimes a posteriori grouping information is used to characterize the different latent classes.

5.4 Applications of Mixture IRT Using Personality Data

Mixture IRT models have been applied to identify classes who systematically differ on some aspects of the personality trait being measured. From a psychological point of view this is very interesting. Groups of persons may differ in the way a personality construct is applicable. Unidimensional IRT models assume that the probability of endorsing an item is similar for all persons in the sample. However, this assumption may not be realistic. For example, consider the use of a multifaceted trait construct. Many trait constructs are composed of content heterogeneous sets of
items. Persons may differ in the endorsement probability of different sets of items because there are differences in trait manifestations between groups of persons. An alternative is then to use a mixture IRT model allowing differences between the parameters between subpopulations. For example, Reise and Gomel (1995) applied mixture IRT to responses on the 14 true-false items of the Positive Interpersonal Engagement scale (PIE). A two-class solution was most interpretable. The first class consisted of Agentic type of persons (i.e., persons who are dominant in social transactions) and the second class of Communal type of persons (i.e., persons who are more sociable and like to have people around). Maij-de Meij, Kelderman, and van der Flier (2005) used mixture IRT to assess qualitative individual differences in self-disclosure patterns. A model with three latent classes was identified. Class 1 consisted of persons with a high-level of self-disclosure but who were most selective to whom they show their self-disclosure, whereas Class 2 consisted of persons with an opposite pattern of Class 1, and persons in Class 3 had medium scores and were not very selective to whom they show their self-disclosure. Thus, in both the Reise and Gomel (1995) and the Maij-de Meij et al. (2005) studies qualitative differences were found.

Mixture IRT models have also been applied to obtain insight into different types of response behavior in personality assessment with both dichotomous (e.g., Reise & Gomel, 1995) and polytomous items (e.g., Austin et al., 2006). Hernández, Drasgow, and González-Romá (2004) applied mixture IRT to examine differences in the way persons respond to an ordered scale with a middle category (the ‘I don’t know’ option). They identified two latent classes, in the first class most persons did not respond to the middle category, whereas persons in the second class selected this category much more often.

Rost, Carstensen, and von Davier (1997) and Austin et al. (2006) investigated whether they could identify different latent classes related to different response scale usage on the NEO-FFI personality questionnaire (Costa & McCrae, 1992). Austin et al. (2006) used all Big Five scales, whereas in the Rost et al. (1997) study only the Conscientiousness and the Extraversion scale were used. Results from the Austin et al. (2006) study yielded a two-class solution for Neuroticism, Extraversion, Agreeableness, and Conscientiousness. Persons in one class showed preferences for the extremes of the response scale, whereas persons in the other class showed preferences for responding in the middle of the scale. For Openness a three-class solution fitted best, however, this three-class solution was difficult to interpret. Austin et al. (2006) attributed this to multidimensionality of the Openness scale. Factor analysis resulted in two factors: one factor was related to intellectual/aesthetic
curiosity and one factor was related to conventionality and acceptance of authority. With respect to the Conscientiousness scale, the results of the Rost et al. (1997) study were similar as in the Austin et al. (2006) study. Rost et al. (1997) found one class with persons that preferred extreme ratings and one class with persons with moderate ratings. The results with regard to the Extraversion were different from the Austin et al. (2006) study. Although a four-class solution fitted best, a two-class solution was easier to interpret. This two-class solution reflected individual differences with regard to two dimensions of Extraversion, namely Sociability and Impulsivity. These later results are in line with the results regarding structural differences.

5.5 Predictive Validity of Trait Estimates Obtained from Different Qualitative Groups

Thus, mixture IRT has been applied in different settings for different purposes and it has shown its usefulness. Mixture IRT takes response behavior into account, which provides the potential for a better estimate of an individual’s latent trait score. However, like Eid and Zickar (2007) suggested “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). Although several authors have suggested that inconsistent response behavior and response tendencies may have an effect on the predictive validity of a measurement instrument, it is difficult to draw an unambiguous conclusion from the literature. For example, Meijer (1997; see also Meijer & Nering, 1997) using simulated data showed that there was a decrease in the validity of test scores when persons showed extreme forms of response styles (i.e., persons answering most items randomly) when the percent of persons in a sample with extreme response styles was 15% or larger and test validity was at least moderate (i.e., $r = .3 - .4$). Removing classes of persons with extreme response styles resulted in modest increases in test validity. However, Meijer (1997) also showed that for the group with extreme response styles test-criterion relations were relatively lower than for those with IRT consistent response patterns. Also Schmitt, Chan, Sacco, McFarland, and Jennings (1999; see also Schmitt, Cortina, & Whitney, 1993) found that for groups that were less in accord with a unidimensional IRT model, test scores were less predictable than those whose response patterns were more consistent with an IRT model. Further analysis showed that the validity was
substantially larger for model-fitting examinees than for non-model-fitting examinees (i.e., \( r = .34 \) versus \( r = .19 \)).

Maij-de Meij, Kelderman, and van der Flier (2008) were one of the first researchers to investigate whether the predictive validity of trait estimates obtained from mixture IRT models was higher compared to the predictive validity of trait estimates obtained from a unidimensional IRT model. They distinguished different classes with different response styles. Their application of mixture IRT to the Extraversion and Neuroticism scale of an often used Dutch personality questionnaire resulted in a better prediction for the Neuroticism scale, but not for the Extraversion scale. As a criterion variable they used the score from a psychologist on a 5-point scale about the extraversion and emotional stability level of the applicant during the selection interview. Although the psychologist did not know the applicant’s test results, the criterion was strongly related to the overall impressions of only one psychologist.

In addition to the studies cited above, the present study will extend the literature by applying mixture IRT to a Conscientiousness scale. The predictive validity of trait estimates obtained from mixture IRT models is compared to trait estimates obtained from a unidimensional IRT model using a 360 degree feedback questionnaire on Conscientiousness as a criterion measure.

5.6 Method

5.6.1 Instruments

Workplace Big Five (WB5)
The WB5 (Schakel, Smid, & Jaganjac, 2007) is a computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. It consists of 144 items, distributed over five scales (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a 5-point Likert scale. The answer most indicative of the trait being measured is scored “5” and the answer least indicative of the trait is scored “1”. For this study, we selected the Conscientiousness scale. This scale consists of 30 items, equally distributed over five subscales (C1 Perfectionism, C2 Organization, C3 Drive, C4 Concentration, and C5 Methodicalness). Perfectionism assesses the degree to which we strive for perfect results, Organization assesses the degree to which we work in an organized, structured manner, Drive assesses the degree to which we always strive to achieve
more, Concentration assesses how we keep on concentrating our attention on a task and Methodicalness assesses the degree to which we plan with foresight and in detail.

The WB5 is a Dutch version of the Workplace Big Five Profile constructed by Howard and Howard (2001). This profile is based on the NEO-PI-R (Costa & McCrae, 1992) and adapted to workplace situations. For the Dutch version, both conceptual analyses and exploratory factor analyses showed the Big Five structure (Schakel et al., 2007).

**Reflector**

The Reflector (Schakel et al., 2007) is a 360 degree feedback questionnaire that can be used to obtain information from different sources concerning a person’s performance on the job. This information often provides the basis for personnel development. The questionnaire is constructed on the basis of selected job-relevant behavioral skills and consists of 215 statements of behavioral skills in different content areas. The questionnaire has to be filled out by the person to be evaluated and by different persons from his or her direct working environment, such as supervisors, colleagues, subordinates, and customers. The items are scored on a 5-point Likert scale. The answer most indicative of the behavioral skill being measured is scored “5” and the answer least indicative of the skill is scored “1”. To evaluate a person, different behavioral skills from different content areas can be selected. In this study, we selected data from persons who were evaluated on the basis of the 15-item Conscientiousness Reflector scale, which consists of 15 Conscientiousness related behavioral skills. This scale was used as a criterion measure. An example of a skill that was evaluated is “Plans time for preparation and unexpected matters”.

### 5.6.2 Participants and Procedure

Data were collected in the context of a career development procedure. Persons in the career development procedure filled out the WB5 and at least three different persons from his or her working environment filled out the Reflector. The sample consists of 9283 persons with a mean age of 40.8 ($SD = 8.71$); 45.8% mostly White men and 47.8% mostly White women; for 3.7% age was unknown and for 6.4% gender was unknown. 16.3% of the participants had a university degree, 44.6% had higher education, and 38.5% secondary education; for .6% educational level was unknown. These persons were occupied in different industries, most of the persons worked in government, financial, or information technology institutions. On average
they had 12.7 years of work experience ($SD = 10.92$) and most of the persons hold middle- or high-level management positions.

For the Reflector scores we selected persons who were evaluated by at least three other persons on the Conscientiousness Reflector scale. Data were available from supervisors, colleagues, subordinates, and customers. Literature on 360 degree feedback shows that raters agree more with each other than they do with the person to be evaluated (e.g., Atwater & Yammarino, 1992; Carless, Mann, & Wearing, 1998; Furnham & Stringfield, 1998). In our sample the numerical values of the correlations of the self-other and other-other ratings were found to be similar to the ones found in the literature, with the exception of the self-supervisor ratings correlation (supervisor-subordinate: $r = .39$, supervisor-colleague: $r = .52$, and colleague-subordinate: $r = .39$, compared to self-supervisor: $r = .32$ and self-colleague: $r = .35$). The ratings of customers did not significantly correlate with the ratings of the other types of raters. Therefore, we decided to exclude data from customers in this study.

Because the other-other ratings correlations were higher than the self-other ratings correlations and because, in general, taking and analyzing each rater type separately yielded too few cases, we first decided to collapse across perspectives. Thus, the Reflector scores were calculated by averaging the responses given by at least three supervisors, colleagues and/or subordinates. For 630 out of 9283 participants the Conscientiousness Reflector scores were available. Each person was evaluated on average by 3.90 others ($SD = 0.96$), more specifically they were evaluated on average by 1.15 supervisors ($SD = 0.58$), 2.55 colleagues ($SD = 1.05$) and .20 subordinates ($SD = 0.62$).

Second, to assess whether a Reflector score based on ratings from one perspective yields similar predictive validity results in comparison with a Reflector score based on ratings from supervisors, colleagues and subordinates, we selected persons who were evaluated by at least two colleagues. For 607 out of 9283 participants the Reflector scores of at least two colleagues were available (for the supervisor perspective $n = 124$ and for the subordinate perspective $n = 32$) and their Reflector scores were calculated by averaging the responses given by at least two colleagues. Each person was evaluated on average by 2.71 colleagues ($SD = 0.89$). Using both Reflector scores from one perspective (colleagues) and different perspectives (colleagues, supervisors, and subordinates) enabled us to investigate the influence of different perspectives as compared to one perspective.

The 630 persons in our study were evaluated by 2458 supervisors, colleagues and subordinates. Each of the 2458 evaluations was treated as a separate one to compute coefficient alpha for the 15 items. Coefficient alpha for the Conscientiousness
Reflector scale equalled .88, whereas coefficient alpha equalled .87 when only evaluations of colleagues were used ($n = 1645$).

### 5.6.3 Analysis

#### Scalability

To explore the quality of the data, we computed classical statistics and conducted a nonparametric IRT analysis using the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0; Molenaar & Sijtsma, 2000). We checked the assumptions of the nonparametric Mokken model of monotone homogeneity (MMH, e.g., Sijtsma & Molenaar, 2002) by inspecting the overall scale coefficient $H$ and the item $i$ coefficient $H_i$. These coefficients are used to assess the scalability of the Conscientiousness scale and its items. The $H$ coefficient is a measure to determine whether a set of items form a scale, that is, whether the items relate to each other and allow an ordering of the persons according to their total score. The larger this coefficient, the better we can order persons according to their total score. Also, larger values of the $H_i$ coefficient imply that an item is better able to differentiate among persons and that an item is more related to the scale.

Increasing values of $H$ and $H_i$ between .30 and 1.00 (maximum) reflect better scalability of sets of items and individual items (for a discussion of these measures see for example, Sijtsma & Molenaar, 2002). Furthermore, weak scalability is obtained if $0.30 \leq H < 0.40$, medium scalability if $0.40 \leq H < 0.50$ and strong scalability if $0.50 \leq H < 1$. The choice of the lower bound reflects the strength of the relation between items that is required. If the $H$ value is between .00 and .30, then the items would not have enough in common to trust the ordering of persons by the total score to accurately reflect an ordering on a meaningful unidimensional latent trait. The rules of thumb play a role similar to the requirements for the reliability coefficient in classical test theory. Many psychologists consider a reliability coefficient value of .90 as a lower bound for important decisions about a person, and values of at least .60 or .70 as a lower bound for valid inferences about groups of persons or low-stakes decisions. Because in this study we analyze career development data with lower stakes than selection data, weak scales will suffice. We used this nonparametric IRT technique in addition to the mixture IRT approach to obtain a first impression of the data quality.
Mixture IRT

To apply a mixture IRT model, first a plausible model is fit that assumes a single, homogeneous population. This baseline analysis is then compared with an alternative model incorporating two (or more) subpopulations of persons. If the alternative model provides a clearly superior fit, subgroups of persons may interpret items differently or respond to items according to different processes. If this is the case, observed scores of persons from the different groups may have different meanings and should be interpreted with care. We used the computer program Latent GOLD 4.5 (Vermunt & Magidson, 2005) to estimate the mixture IRT version of the graded response model (GRM; Samejima, 1969, 1997). The GRM is a 2-parameter-logistic (2PL) model for polytomous data. Each item $i$ is described by one item slope parameter or discrimination parameter ($\alpha$ parameter) and two or more location or “difficulty” parameters ($\beta_m$ parameters); the number of location parameters per item is equal to the number of response categories minus 1. The location parameter $\beta_m$ can be interpreted as the point at the latent trait continuum where there is a 50% chance of scoring in category $m$ or higher. The $\alpha$ and $\beta_m$ parameters are used to determine the probability of an examinee to respond in a particular response category (Embretson & Reise, 2000).

Specifically, the mixture version of the GRM was fit by systematically increasing the number of latent classes. Mixture IRT models identify different latent classes with different item parameters. When applying the mixture version of the GRM, classes can have different discrimination parameters and different difficulty parameters or classes can have equal discrimination parameters but have different difficulty parameters. We tested the following two models. Model 1 allows the discrimination and difficulty parameters to differ within each latent class. Model 2 allows the discrimination parameters to differ within a class, but the discrimination parameters for an individual item should be equal across classes. The difficulty parameters in Model 2 may differ within and across classes. Note that a mixture IRT model with one class is the same as a unidimensional IRT model.

Different methods are available to assess model fit and to select the appropriate number of classes. We used the Bayesian information criterion (BIC; Schwarz, 1978) to choose between Model 1 and Model 2. As opposed to the Akaike information criterion, the BIC (Schwarz, 1978) takes the sample size into account, avoiding overparameterization (McLachlan & Peel, 2000). Lower BIC values indicate a better fit and BIC values can be compared across the two models. To select the appropriate number of classes for the chosen model, we used BIC values, the bootstrap $p$-value for the $L^2$-statistic (Vermunt & Magidson, 2005) and the
Conscientiousness in the Workplace: Applying Mixture IRT

misclassification rate. Because of a large number of sparse data (i.e., there are far less subjects than the maximum number of possible response patterns), we estimated the bootstrap $p$-value for the $L^2$-statistic. This method, implemented in LatentGOLD 4.5, generates 300 bootstrap samples based on the parameters of the (chosen) model and re-estimates the model with each bootstrap sample. The $p$-value indicates the amount of bootstrap samples with a poorer fit than the original sample. Generally, the best fitting model has a $p$-value greater than .05 and has the fewest number of parameters (Vermunt & Magidson, 2005). LatentGOLD 4.5 calculates the misclassification rate. Persons are assigned to the class with the highest class membership probability which may lead to misclassification. Besides using these statistical methods, we also checked whether the results could be meaningfully interpreted.

A general problem for mixture IRT models is that local maxima may be found besides the global maximum likelihood solution. Therefore, Latent GOLD 4.5 uses an estimation procedure with multiple sets of random starting values. We conducted our analyses by running 50 startsets and 250 iterations per set. This was also done to avoid convergence problems.

5.7 Results

5.7.1 Descriptive Statistics and Scalability

Item means, standard deviations, and coefficient alpha for the Conscientiousness scale and its subscales are reported in Table 5.1. A first observation is that, in general, the items are relatively “easy” or popular. That is, most respondents chose Categories 4 and 5 indicating that most persons consider themselves as relatively

| Table 5.1 |
| Scale characteristics for the Conscientiousness scale and subscales. |
| # Items | $\alpha$ | $M_{item}$ | $SD_{item}$ | Skewness | Kurtosis |
| Conscientiousness | 30 | .89 | 3.86 | .48 | -.35 | .00 |
| C1 Perfectionism | 6 | .80 | 3.51 | .77 | -.26 | -.45 |
| C2 Organization | 6 | .81 | 4.12 | .68 | -.77 | .20 |
| C3 Drive | 6 | .68 | 4.00 | .58 | -.52 | .10 |
| C4 Concentration | 6 | .74 | 3.57 | .70 | -.34 | -.13 |
| C5 Methodicalness | 6 | .73 | 4.11 | .54 | -.67 | .68 |

$a$ = coefficient alpha.
high on the different subscales of the Conscientiousness scale. Coefficient alpha of the subscale Drive was low relative to the other subscales. Inspection of the item information curves showed flat curves for the items of the subscale Drive, which implies that these items do not allow reliable measurement. Also, \( H_i \) values of the Drive items within the total Conscientiousness scale were low compared to the values of the other scales (most Drive items had \( H_i \) values around .20, while most items of the other subscales had \( H_i \) values around .25) and the overall \( H \) value for the Conscientiousness scale increased from .25 to .29 when removing the items of the subscale Drive. Moreover, inspection of the content of the items clearly showed that these items are sensitive to social desirable responding. Items that were included in this scale were, for example, “avoids new responsibilities” (reverse-scored) and “always wants to perform to the best of his/her ability”. Therefore, we removed the subscale Drive from the scale. Furthermore, we removed Item 4 from the

<table>
<thead>
<tr>
<th>Item</th>
<th>Subscale</th>
<th>Total</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item1</td>
<td>C1 Perfectionism</td>
<td>.35</td>
<td>.36</td>
<td>.34</td>
<td>.28</td>
<td>.30</td>
</tr>
<tr>
<td>Item2</td>
<td>C1 Perfectionism</td>
<td>.24</td>
<td>.24</td>
<td>.22</td>
<td>.19</td>
<td>.15</td>
</tr>
<tr>
<td>Item3</td>
<td>C1 Perfectionism</td>
<td>.29</td>
<td>.27</td>
<td>.24</td>
<td>.23</td>
<td>.21</td>
</tr>
<tr>
<td>Item5</td>
<td>C1 Perfectionism</td>
<td>.36</td>
<td>.37</td>
<td>.29</td>
<td>.29</td>
<td>.28</td>
</tr>
<tr>
<td>Item6</td>
<td>C1 Perfectionism</td>
<td>.31</td>
<td>.34</td>
<td>.28</td>
<td>.24</td>
<td>.19</td>
</tr>
<tr>
<td>Item7</td>
<td>C2 Organization</td>
<td>.32</td>
<td>.33</td>
<td>.31</td>
<td>.27</td>
<td>.31</td>
</tr>
<tr>
<td>Item8</td>
<td>C2 Organization</td>
<td>.31</td>
<td>.34</td>
<td>.33</td>
<td>.24</td>
<td>.31</td>
</tr>
<tr>
<td>Item9</td>
<td>C2 Organization</td>
<td>.34</td>
<td>.35</td>
<td>.35</td>
<td>.31</td>
<td>.32</td>
</tr>
<tr>
<td>Item10</td>
<td>C2 Organization</td>
<td>.38</td>
<td>.38</td>
<td>.38</td>
<td>.31</td>
<td>.35</td>
</tr>
<tr>
<td>Item11</td>
<td>C2 Organization</td>
<td>.39</td>
<td>.40</td>
<td>.37</td>
<td>.34</td>
<td>.38</td>
</tr>
<tr>
<td>Item12</td>
<td>C2 Organization</td>
<td>.39</td>
<td>.38</td>
<td>.37</td>
<td>.33</td>
<td>.38</td>
</tr>
<tr>
<td>Item19</td>
<td>C4 Concentration</td>
<td>.26</td>
<td>.26</td>
<td>.24</td>
<td>.21</td>
<td>.25</td>
</tr>
<tr>
<td>Item20</td>
<td>C4 Concentration</td>
<td>.31</td>
<td>.29</td>
<td>.29</td>
<td>.18</td>
<td>.25</td>
</tr>
<tr>
<td>Item21</td>
<td>C4 Concentration</td>
<td>.32</td>
<td>.27</td>
<td>.26</td>
<td>.17</td>
<td>.29</td>
</tr>
<tr>
<td>Item22</td>
<td>C4 Concentration</td>
<td>.33</td>
<td>.31</td>
<td>.32</td>
<td>.18</td>
<td>.22</td>
</tr>
<tr>
<td>Item23</td>
<td>C4 Concentration</td>
<td>.31</td>
<td>.31</td>
<td>.32</td>
<td>.22</td>
<td>.19</td>
</tr>
<tr>
<td>Item24</td>
<td>C4 Concentration</td>
<td>.24</td>
<td>.24</td>
<td>.24</td>
<td>.18</td>
<td>.24</td>
</tr>
<tr>
<td>Item25</td>
<td>C5 Methodicalness</td>
<td>.22</td>
<td>.22</td>
<td>.24</td>
<td>.17</td>
<td>.21</td>
</tr>
<tr>
<td>Item26</td>
<td>C5 Methodicalness</td>
<td>.34</td>
<td>.34</td>
<td>.33</td>
<td>.29</td>
<td>.34</td>
</tr>
<tr>
<td>Item28</td>
<td>C5 Methodicalness</td>
<td>.37</td>
<td>.37</td>
<td>.36</td>
<td>.31</td>
<td>.36</td>
</tr>
<tr>
<td>Item29</td>
<td>C5 Methodicalness</td>
<td>.29</td>
<td>.27</td>
<td>.28</td>
<td>.25</td>
<td>.28</td>
</tr>
<tr>
<td>Item30</td>
<td>C5 Methodicalness</td>
<td>.36</td>
<td>.35</td>
<td>.36</td>
<td>.29</td>
<td>.34</td>
</tr>
</tbody>
</table>

\[
H = .32 \quad .32 \quad .31 \quad .25 \quad .28
\]
Conscientiousness in the Workplace: Applying Mixture IRT

Perfectionism subscale and Item 27 from the Methodicalness subscale for subsequent analyses because of very low $H_i$ values. As a result, the overall $H$ coefficient increased ($H = .29$ became $H = .31$).

From inspecting the category response functions and option endorsement proportions it was clear that Categories 1 and 2 had very low endorsement proportions. Therefore, we collapsed these two categories. This did not affect mixture IRT results. Table 5.2 (third column) shows the $H$ and $H_i$ coefficients for the total Conscientiousness scale with four response categories. All items covaried positively and although there were a few items with $H_i \leq .25$ we did not remove these items because it did not have an effect on the overall $H$ coefficient. Note that all scales are weak scales. In practice, weak scales will almost always result when scales consists of items that have some content heterogeneity (see e.g., Reise & Waller, 2009).

For these 22 items within the four subscales, in Table 5.3 the correlations between the subscales and the correlations between the final Conscientiousness scale and the subscales are given. From this table it is clear that the subscales are moderately correlated with each other and are highly correlated with Conscientiousness.

Table 5.3
Correlations between the subscales and between the final Conscientiousness scale and the subscales.

<table>
<thead>
<tr>
<th></th>
<th>C1 Perf</th>
<th>C2 Orga</th>
<th>C4 Conc</th>
<th>C5 Meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2 Organization</td>
<td>.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4 Concentration</td>
<td>.43</td>
<td>.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5 Methodicalness</td>
<td>.41</td>
<td>.66</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.75</td>
<td>.85</td>
<td>.80</td>
<td>.78</td>
</tr>
</tbody>
</table>

Note. Perf = Perfectionism; Orga = Organization; Conc = Concentration; Meth = Methodicalness.

5.7.2 Mixture IRT

Before we conducted the mixture IRT analyses for Model 1 and Model 2, we checked whether the assumption of a single homogenous population holds for the data\(^1\). Therefore, we tested a model which assumes equal discrimination and

\(^1\) A violation of the assumption of a homogenous population means that the population consists of subpopulations that significantly differ from each other on their mean latent trait score. When this is the case, the data should be analyzed as multiple group. If a mixture IRT model is used without checking this assumption, it may be difficult to interpret the results because differences from the mixture IRT analyses may be due to existing differences in mean latent trait score between groups.
difficulty parameters across classes, but allows for differences in mean and standard deviation of the latent trait scores across classes. Results showed that the assumption of a homogenous population holds for our data.

Table 5.4 displays the fit statistics for the mixture version of the GRM for the Conscientiousness scale\(^2\). Model 1 allows the discrimination and difficulty parameters to differ within each latent class. Model 2 assumes equal discrimination parameters for each individual item across all latent classes and allows the difficulty parameters to differ within each latent class. Comparing the BIC values of both models showed lower BIC values for Model 2. Therefore, we used Model 2 for further analyses. When only using BIC values to select the appropriate number of classes, the 12-class solution from Model 2 should be selected. However, the classes become too small to be interpretable. Column 5 in Table 5.4 displays the bootstrap \(p\)-values for the \(L^2\)-statistic. According to this statistic, the two-class solution should be selected (i.e., \(p > .05\) and fewest parameters). However, the four-class solution might also be worth considering (\(p = .41\)). From column 6 in Table 5.4 it is clear that by adding a class the misclassification rate increases. Therefore, we investigated the

<table>
<thead>
<tr>
<th># m</th>
<th>log (l)</th>
<th>npar</th>
<th>BIC (log (l))</th>
<th>bootstrap (p)-value</th>
<th>misclass. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-220905.37</td>
<td>88</td>
<td>442614.71</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>2</td>
<td>-215869.50</td>
<td>177</td>
<td>433356.06</td>
<td>34 .091</td>
<td>.091</td>
</tr>
<tr>
<td>3</td>
<td>-212923.19</td>
<td>266</td>
<td>428276.54</td>
<td>60 .117</td>
<td>.117</td>
</tr>
<tr>
<td>4</td>
<td>-211793.90</td>
<td>355</td>
<td>426831.05</td>
<td>41 .136</td>
<td>.136</td>
</tr>
<tr>
<td>5</td>
<td>-210774.80</td>
<td>444</td>
<td>425605.95</td>
<td>51 .174</td>
<td>.174</td>
</tr>
<tr>
<td>6</td>
<td>-209984.98</td>
<td>533</td>
<td>424839.41</td>
<td>74 .189</td>
<td>.189</td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-220905.37</td>
<td>88</td>
<td>442614.71</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>2</td>
<td>-216021.41</td>
<td>155</td>
<td>433458.88</td>
<td>34 .091</td>
<td>.091</td>
</tr>
<tr>
<td>3</td>
<td>-213031.64</td>
<td>222</td>
<td>428091.46</td>
<td>60 .117</td>
<td>.117</td>
</tr>
<tr>
<td>4</td>
<td>-211960.48</td>
<td>289</td>
<td>426561.25</td>
<td>41 .136</td>
<td>.136</td>
</tr>
<tr>
<td>5</td>
<td>-211032.64</td>
<td>356</td>
<td>425317.68</td>
<td>51 .174</td>
<td>.174</td>
</tr>
<tr>
<td>6</td>
<td>-210216.10</td>
<td>423</td>
<td>424296.71</td>
<td>74 .189</td>
<td>.189</td>
</tr>
</tbody>
</table>

Note. \# m = number of latent classes; log \(l\) = log-likelihood statistic; npar = number of parameters; BIC = Bayesian information criterion; misclass. rate = misclassification rate; n.a. = not applicable.

\(^2\) We also applied the mixture version of the generalized Partial Credit Model (Muraki, 1992). However, the mixture version of the GRM fitted better; it provided lower BIC values.
results of the two-class solution, the three-class solution, and the four-class solution. Based on its bootstrap $p$-value for the $L^2$-statistic ($p = .51$), we also investigated the results of the five-class solution. The four-class solution yielded the most interpretable and meaningful results. In the four-class solution 35.8%, 28.4%, 18.2%, and 17.7% of the persons belong to Class 1, Class 2, Class 3, and Class 4, respectively. The mean class membership probabilities were $p = .88$, $p = .87$, $p = .84$ and $p = .84$ for these classes. For only 3.6% of the 9283 persons $p < .50$ and for 11.3% of the persons $p < .60$.

**Interpretation of the Classes**

To investigate the response scale usage of the persons in the different classes, we plotted the distribution of the response categories for the whole Conscientiousness scale. Figure 5.1 shows that, as expected, Categories 3 and 4 were popular across

![Figure 5.1: Distribution of the chosen response categories for the four-class solution](image)

classes. Differences in response scale usage are more pronounced when we inspect the response frequencies at the item level. In Figure 5.2 we report the item category response frequencies for each class in the four-class solution. From Figure 5.2 it is clear that persons in the Classes 1 and 3 are most consistent in their response scale usage. Persons in Class 1 prefer Category 3 most of the time, whereas persons in Class 3 prefer Category 4 most of the time. However, consider the response scale usage of the persons in Class 2 and Class 4 on the subscales Perfectionism and Concentration, respectively. Persons in Class 2 prefer Category 1 most of the time
Figure 5.2: Distribution of the chosen response categories per item for Class 1 (upper left panel), Class 2 (upper right panel), Class 3 (lower left panel), and Class 4 (lower right panel).
for the Perfectionism items (Items 1-6), whereas persons in Class 4 prefer Category 1 for the Concentration items (Items 19-24). The responses on the two subscales Organization (Items 7-12) and Methodicalness (Items 25-30) are similar for the four classes; Categories 3 and 4 have the highest probability of being selected.

Thus, in the four-class solution persons in the different classes differ in particular with respect to how they respond to items from the subscales Perfectionism and Concentration. Table 5.5 displays the mean total scores on the Conscientiousness scale and its subscales for each latent class. Persons in Class 2 have lower scores on Perfectionism compared to persons in Classes 1, 3, and 4, whereas persons in Class 4 have lower scores on Concentration compared to persons in Classes 1, 2, and 3. For Perfectionism and Concentration, the differences in the mean total scores between the classes are often more than one standard deviation. Calculating the effect sizes showed that the differences between the classes can be labeled as ‘large’ for Perfectionism and Concentration. Cohen’s $d$ ranged from 1.11 to 2.63 for the differences in mean total score on Perfectionism (with the exception of the small difference between Classes 3 and 4, Cohen’s $d = 0.33$) and from 1.18 to 3.01 for the differences between the classes on Concentration (with the exception of the small difference between Classes 1 and 2, Cohen’s $d = 0.19$). The differences in mean total scores for the other subscales are smaller, therefore, the effect sizes are much smaller.

Table 5.5

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>C1 Perf</td>
<td>12.78</td>
<td>2.76</td>
<td>10.35</td>
<td>3.38</td>
</tr>
<tr>
<td>C2 Orga</td>
<td>17.87</td>
<td>3.29</td>
<td>19.31</td>
<td>4.03</td>
</tr>
<tr>
<td>C4 Conc</td>
<td>15.18</td>
<td>2.88</td>
<td>15.65</td>
<td>3.92</td>
</tr>
<tr>
<td>C5 Meth</td>
<td>15.02</td>
<td>2.17</td>
<td>16.66</td>
<td>2.73</td>
</tr>
<tr>
<td>Conscien</td>
<td>60.85</td>
<td>9.17</td>
<td>61.97</td>
<td>11.71</td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
</tr>
</tbody>
</table>

Note. Perf = Perfectionism; Orga = Organization; Conc = Concentration; Meth = Methodicalness; Conscien = Conscientiousness.

To further explore these differences we were curious to know whether the scalability between classes was different, that is whether Conscientiousness may be a rather homogeneous construct for some groups of persons, whereas it is a more heterogeneous construct for others.
5.7.3 Scalability Within Classes

To investigate the scalability for the different classes, we conducted a Mokken scale analysis and we compared scalability results for the whole sample with results obtained for the different classes. Note that the comparison is made between scalability results for each class and the whole sample that includes that class. Results are shown in Table 5.2 (last four columns). In general, the overall scalability remained the same for Class 1 (i.e., $H = .32$), whereas it decreases for Classes 2, 3, and 4 (i.e., $H = .31$, $H = .25$ and $H = .28$ for Classes 2, 3, and 4, respectively). Interesting is to consider the individual $H_i$ coefficients. For Class 1, the $H_i$ coefficients are similar to the $H_i$ coefficients for the whole sample. For Class 2, the $H_i$ coefficients of the Perfectionism items are lower compared to the $H_i$ coefficients for the whole sample, the average absolute differences in $H_i$ coefficients is .04 for those items. For Class 4, the $H_i$ coefficients for both the Perfectionism and Concentration items are lower, the average absolute differences in $H_i$ coefficients are .08 and .06, respectively. Furthermore, for Class 3, all $H_i$ coefficients are lower, the average absolute differences in $H_i$ coefficients between Class 3 and the whole sample are .06, .06, .11, and .05 for the Perfectionism, Organization, Concentration and Methodicalness items, respectively. Because $H$ coefficients are related to the discrimination parameters in parametric IRT models\textsuperscript{3}, these results thus imply that Perfectionism and Concentration are less related to Conscientiousness for persons in the Classes 2 and 4 than for persons in Class 1. The whole Conscientiousness scale is less scalable for persons in Class 3.

5.7.4 Psychological Meaning of Class Differences

To further characterize the different classes, we first took the demographic variables gender, education, and age into account. Figure 5.3 shows that there are more males than females in the Classes 1 and 2, and more females than males in the Classes 3 and 4. Furthermore, in Class 2 there are more persons with a university degree or higher education. Persons in Class 3 had more often a secondary educational level. There were only small differences in age between the classes.

\textsuperscript{3} Although $H_i$ coefficients and $\alpha$ parameters both indicate the slope of the item response function, both statistics are also sensitive to different characteristics of the data. $H_i$ is strongly influenced by the probability distribution of the latent trait values which also prevents a researcher from using a scale in a population where it cannot discriminate between persons. In the literature there is a strong emphasize on selecting items with $H_i$ values larger than some lower bound as, say, $H_i = .3$. 
Second, we tried to characterize the persons in the different classes using scale scores from other personality scales of the WB5. Based on the literature it was difficult to interpret the results regarding Concentration, but for Perfectionism some interesting observations could be made. Literature on different forms of Perfectionism (e.g., Stoeber & Otto, 2006) usually differentiates three groups: healthy perfectionists (high score on Perfectionism, low score on Neuroticism), unhealthy perfectionists (high score on both Perfectionism and Neuroticism) and non-perfectionists (low score on both Perfectionism and Neuroticism). This distinction is based on the two forms of Perfectionism suggested by Hamachek (1978), a positive form called ‘normal perfectionism’ (persons who enjoy pursuing their perfectionistic strivings) and a negative form called ‘neurotic perfectionism’ (persons who suffer from their perfectionistic strivings). Our results showed three levels of Perfectionism; the mean Perfectionism score was lowest for persons in Class 2, persons in Class 1 had relatively moderate Perfectionism scores and persons in Classes 3 and 4 had the highest Perfectionism scores (see Table 5.5 for the mean total scores on the subscale Perfectionism for each class).

To further explore the differences in answering behavior across the four groups and to characterize the persons in each class, we used ancillary information from the Neuroticism scale because the distinction between positive and negative Perfectionism is related to Neuroticism. We calculated the mean total scores for each class on the Neuroticism scale and found mean Neuroticism scores of 73.7 (SD
= 13.27), 65.9 (SD = 14.65), 67.1 (SD = 16.22) and 78.8 (SD = 16.92) for Class 1, Class 2, Class 3, and Class 4, respectively. Thus, persons in Class 4 have the highest mean scores on Neuroticism, whereas persons in Classes 2 and 3 have the lowest mean scores on Neuroticism. Furthermore, persons in Class 2 may have some characteristics of ‘non-perfectionists’ (relatively low score on both Perfectionism and Neuroticism), persons in Class 4 of ‘unhealthy perfectionists’ (relatively high scores on both Perfectionism and Neuroticism) and persons in Class 3 of ‘healthy perfectionists’ (high score on Perfectionism, low score on Neuroticism). Persons in Class 1 (the largest group) had moderate scores on both Perfectionism and Neuroticism. These persons could be labeled as ‘normal perfectionists’.

Also remember that there are more females than males in Class 4, which may also be an explanation of the higher Neuroticism scores, since personality research consistently reports that women score higher on Neuroticism than men (e.g., Lynn & Martin, 1997).

5.7.5 Predictive Validity

Using Data from at least Three Supervisors, Colleagues and/or Subordinates

To evaluate the predictive validity of the trait estimates obtained from the mixture IRT model compared to these obtained from a unidimensional IRT model, the estimated trait scores under each model were related to the Reflector scores. There were \( n = 630 \) persons for whom criterion data were available; \( n = 213, n = 114, n = 184 \) and \( n = 119 \) for Classes 1, 2, 3, and 4, respectively. Because different models hold within the classes, the estimated trait scores are not on the same trait continuum and cannot be compared for all classes together. Therefore, the relation between the trait scores and the external criterion measure was evaluated for each latent class separately. The correlation between the trait estimates obtained from the unidimensional IRT model and the Reflector scores equalled \( r = .35 \). The correlations between the trait estimates obtained from the mixture IRT model and the external criterion were \( r = .41, r = .40, r = .32 \) and \( r = .45 \) for Class 1, Class 2, Class 3, and Class 4, respectively.

Although differences are sometimes small, there seems to be a trend that the predictive validity of the trait estimates in Classes 1, 2, and 4 increases as compared to the predictive validity of the trait estimates obtained from the unidimensional IRT model, whereas it decreases for the trait estimates in Class 3. The response patterns of the persons in Class 3 were less scalable than the patterns in the other classes.
which resulted in lower test-criterion relations. These results are in line with the simulation results found in Meijer (1997) using person-fit statistics.

**Using Data from at least Two Colleagues**

There were \( n = 607 \) persons for whom criterion data were available; \( n = 207, n = 105, n = 180 \) and \( n = 115 \) in Classes 1, 2, 3 and 4, respectively. The correlation between the trait estimates obtained from the unidimensional IRT model and the Reflectors scores equalled \( r = .30 \). The correlations between the trait estimates obtained from the mixture IRT model and the external criterion were \( r = .38, r = .38, r = .24 \) and \( r = .43 \) for Class 1, Class 2, Class 3, and Class 4, respectively. These results are in line with the results reported above. The predictive validity of the trait estimates in Classes 1, 2, and 4 increased, whereas it decreased for the trait estimates in Class 3 when using mixture IRT as compared to unidimensional IRT.

**5.8 Discussion**

Reise and Gomel (1995) suggested with respect to the use of mixture IRT models that “explorations will uncover many situations where the largest qualitative differences fall along psychological lines as opposed to racial, gender, or age” (p. 354). Analyzing empirical data from a Conscientiousness scale, we found some evidence that this is the case. Persons in the different classes differed in response scale usage on the subscales Perfectionism and Concentration, indicating structural differences between the classes. Further inspection of the scalability of the Conscientiousness scale for each class separately showed that Conscientiousness is a less scalable and a more heterogeneous construct for persons in the Classes 2 and 4 than for persons in Class 1.

Structural differences between groups with respect to response scale usage are in line with results obtained by Reise and Gomel (1995) and Maij-de Meij et al. (2005), who both used mixture IRT to assess qualitative individual differences between groups of persons. From a psychological point of view, this is most interesting because these analyses reflect the fact that the trait structure of Conscientiousness may be qualitatively different for different groups. Based on our four-class solution and the structural differences, we tried to find a psychological meaning for these group differences. Although our interpretation was post hoc, we showed that there were differences in the way persons interpreted the Conscientiousness scale. Future research may corroborate these findings by, for example, comparing the characteristics of different classes of persons with their scores on Perfectionism
scales like the Revised Almost Perfect Scale (APS-R; Slaney, Rice, Mobley, Trippi, & Ashby, 2001) and may also take recent results from Roberts et al. (2005) about the lower-order structure of Conscientiousness into account.

We further observed that there is a trend that mixture IRT modeling lead to stronger test-criterion relations for three out of the four classes. Maij-de Meij et al. (2008) concluded that mixture IRT models provide possibilities to improve the prediction of external criteria, but that this may vary across scales. They found stronger test-criterion relations for the Neuroticism scale and weaker relations for the Extraversion scale. Our findings are in line with their results on the Neuroticism scale.

How can these results help to improve the assessment practice? An advantage of using mixture IRT models is that they may provide a better insight in the way persons (ranging from psychiatric patients to job applicants) fill out a questionnaire. Reducing heterogeneity has important implications for research and practice. For example in career development, it provides the opportunity to focus on specific subpopulations, both to identify more specific career strategies and to investigate stronger relations to personality characteristics with reduced misclassification. Or in a clinical setting, where focusing on specific subpopulations may ease the use of more specific treatments for different subpopulations (e.g., Nestadt et al., 2009). Thus, mixture IRT models can be used to identify subtypes on personality instruments.

Our results suggest that one should be careful in interpreting the Conscientiousness scores and our results are in line with the call for a more detailed approach to personality assessment (e.g., Hough & Oswald, 2000; Paunonen et al., 2003; Roberts et al., 2005). To further illustrate this, consider Person 3704 (Class 1, sum score 66), Person 6640 (Class 2, sum score 64) and Person 7888 (Class 4, sum score 64). All these persons have a high class membership probability of belonging to that class. According to their sum score, an assessor may be tempted to conclude that these persons have similar Conscientiousness levels. However, taking a closer look at the sum scores on the subscales (see Figure 5.4), it is questionable whether the similar Conscientiousness scores have the same meaning for these three persons and adapting their trait scores may be a better option.
In practice, however, adapting test scores through mixture IRT may not always be allowed. For example, in the US, The Civil Rights Act of 1964 prevents score adjustments based on ethnicity (i.e., on the basis of manifest groups), and adjusting scores on the basis of latent groupings may also be problematic. This may be less problematic in low-stakes settings, such as career development, but more problematic in high-stakes settings, such as in personnel selection.

With respect to the applicability of the trait constructs, our results echo the discussion raised by Bem and Allen (1974) that some trait dimensions apply to some people some of the time. Our analyses showed that different measurement models are needed for different groups of persons. Furthermore, many personality constructs are multifaceted, that is, composed of content heterogeneous sets of items. Much research investigated the best way to measure multifaceted concepts. The question is then whether researchers score the whole scale or score the content subscales. One of the problems of scoring subscales is that these scales are relatively short and unreliable. When subscales are highly intercorrelated the bifactor model may be used to investigate the common and specific variance (e.g., Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). An alternative is to use mixture IRT modeling. Although it probably depends on the application envisaged, mixture IRT modeling provides an alternative to the multidimensional solution to complex trait constructs.
Abstract
When using psychological tests in practice the detection of invalid protocols is important. Within the realm of modern test theory, new approaches have been developed. The potential usefulness of these new approaches is discussed and their relation with existing methods is explored. An empirical example shows that different approaches are sensitive to different types of invalid response behavior and that the usefulness of different approaches depends on the characteristics of the data. Finally, we sketch the practical implications of the use of different validity indicators in personality assessment.

This chapter has been submitted for publication as:
6.1 Introduction

Psychological tests and inventories play an important role in psychological practice, contributing to many decisions that shape individual’s upbringing, education, and careers. Test results assist in identifying talented individuals or in diagnosing individuals who need clinical treatment. Because tests affect people’s lives, they should be constructed according to the highest quality standards. The requirements with respect to the objectivity, reliability, and validity of psychological assessment are increasing and the adaptation of strong psychometric methods is strongly encouraged in different fields of psychology and health research (e.g., Reeve et al., 2007).

An important part of psychological assessment is personality assessment. Personality measures can predict a variety of important outcome variables, such as physical and psychological health, social functioning, and job performance (Ozer & Benet-Martínez, 2006, for an excellent review). In personality assessment often self-report inventories are used to map personality characteristics. The use of self-report inventories is, however, not without problems. Several authors discussed the lack of high quality items, the presence of many similar questions, and the existence of long inventories (sometimes consisting of 200-400 items, e.g., Reise & Henson, 2003), so that all kinds of unintended effects occur like motivation problems and fatigue. Low motivation may also be the result when personality measures are administered for basic research purposes. Persons sometimes show little interest in the accuracy of their data and unsystematic and inconsistent responses may result.

Another often encountered problem that may invalidate the responses to personality questions is faking. Research has shown that persons are able to significantly distort their answers on a wide variety of personality measures. Persons who are instructed to present themselves favorably on personality items, or to exaggerate clinical symptoms are able to do so, and can inflate or underestimate the purported level of pathology (e.g., Bagby, Nicholson, Bacchiochi, Ryder, & Bury, 2002). Also, in high-stakes testing settings, such as in personnel selection, it is often very clear which answers will lead to high scores in the preferred direction, and invalid protocols will result (e.g., Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). Although alternatives have been proposed to deal with these invalid protocols such as the forced-choice item format, there are not many existing scales using these types of items.

In this article, we concentrate on the detection of invalid protocols. As a result of motivation problems, faking, social desirability, and other types of response bias, the
detection of invalid protocols has always been a central concern in both the industrial and organizational practice (e.g., personnel selection) as well as the clinical practice (Ben-Porath & Waller, 1992; Tellegen, 1988). Clinicians diagnose and treat individual clients, and as a result they are interested in the validity of individual scale scores. Mean scores and correlations between variables do not alter as a result of a number of unmotivated test takers. However, a forensic or clinical psychologist trying to obtain a picture of a defendant, who is trying to exaggerate his mental illness, and as a result fakes answers to a personality questionnaire, might make serious errors in the diagnosis when relying on the resulting item scores. Therefore, clinicians often use different types of indicators that give information about the validity of individual test scores. Also in personnel selection, the detection of invalid protocols (faking good) is important, since response distortion can reduce the utility of personality measures in job selection (e.g., Rosse, Stecher, Miller, & Levin, 1998; Winkelspecht, Lewis, & Thomas, 2006).

In the literature a plethora of different types of validity indicators have been proposed, partly originating from different fields of psychology and based on different assumptions about the data. For example, new methods have been developed in the context of modern test theory that claim to be sensitive to different types of invalid protocols (e.g., Meijer & Sijtsma, 2001; Egberink, Meijer, Veldkamp, Schakel, & Smid, 2010). However, much research has focused on the development of validity indicators within a well-defined tradition. For example, validity scales are typically developed within the personality assessment area, whereas so called person-fit indices (to be discussed below) have been developed within the psychometric research. Consequently, for a practitioner it is unclear, which method can best be used to investigate specific types of invalid answering behavior.

Therefore, in this article we (1) discuss different methods that have been proposed in different traditions of psychology to detect invalid response protocols, and (2) explore theoretically and by means of empirical data in a personnel selection and career development context which methods can best be used to detect different types of invalid protocols.

### 6.2 Different Types of Response Validity Indicators

Response validity indicators are measures designed to identify persons who are not measured well by a particular scale or measurement instrument. Although different approaches have been proposed, three different basic approaches can be distinguished.
A first approach is to use validity scales. A validity scale attempts to identify a specific form of response inconsistency. For example, the variable response inconsistency (VRIN) scale consists of matched pairs of items that have similar content. Each time the pairs are marked in opposite directions, a point is scored on the scale. The true response inconsistency (TRIN) scale consists of item pairs that are semantically similar if keyed in opposite directions. This scale measures acquiescence, which is the tendency to agree or mark “true” regardless of content. TRIN and VRIN response inconsistency measures have been built into the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008). Other examples of validity scales are social desirability scales, lie scales, and impression management scales.

In personality assessment a tradition exists to detect invalid test scores using different types of validity scales such as the VRIN and the TRIN of the MMPI-2-RF. Research suggests that the utility of validity scales to detect ‘faking bad’ or exaggerating symptomatics is well supported in clinical and forensic psychological practice (for a review see, Nelson, Hoelzle, Sweet, Arbisi, & Demakis, 2010). For example, Pinsoneault (2007) found that different MMPI validity scales had enough power to be used in practice. However in other fields, such as Industrial and Organizational psychology, the usefulness of these scales to detect ‘faking good’ or social desirability is not undisputed. One of the problems of validity scales is that they are known to be confounded with valid personality trait variance, and show a relationship with other content scales (for a review, see, Uziel, 2010). For example, Piedmont, McCrae, Riemann, and Angleitner (2000), and Tett and Christiansen (2007) discussed and reviewed the utility of different types of validity scales and concluded that validity scales were not very useful to detect invalid protocols.

A second approach is the use of model-based response consistency indices or person-fit statistics. In the psychometric and personality literature (e.g., Meijer & Sijtsma, 2001; Reise & Waller, 1993), it has been suggested that invalid test scores can be identified through studying the configuration of individual item scores by means of these person-fit statistics that are proposed in the context of item response theory (IRT, Embretson & Reise, 2000). In IRT, a model is specified that describes the probability of endorsing a particular statement such as “I am often down in the dumps” as a function of a latent trait score (in this case Negative Affectivity). On the basis of an IRT model observed and expected item scores can be compared and many unexpected item scores alert the researcher that the total score may not adequately reflect the trait being measured. Besides the identification of poorly measured individuals, it has also been suggested that these statistics can be used to
increase predictive validity by selecting poorly measured individuals and to assess the fit between an examinee’s personality and a trait construct (i.e., as traitedness indicators, see Reise & Waller, 1993).

A third approach uses statistics to detect different types of response bias. These statistics vary in complexity (see Zijlstra, 2009, for recent developments). For example, in attitude research extreme responding is often defined as the person’s proportional use of the extreme response categories (e.g., corresponding to scores 1 or 5, on 5-point Likert scale), which may vary between zero (if zero responses are in the extreme response category) and one (when all responses are in the extreme response categories). Or midpoint responding as the person’s proportional use of the middle category (i.e., corresponding to scores 3). A potential drawback of these statistics is that persons with a trait score in the middle of the distribution have a high probability of obtaining many responses in the middle category.

Although, these three basic approaches have been discussed in the literature and validity scales are often used in clinical practice, person-fit indices are less popular in practical applications. Therefore, we discuss the usefulness of these indices in more detail.

### 6.3 The Usefulness of Person-fit Indices in Personality Measurement

It has been suggested that person-fit statistics can be of help to identify different types of invalid test protocols, such as random response behavior or social desirable response behavior (e.g., Meijer & Sijtsma, 2001). However, one should be careful in drawing these conclusions. For example, the usefulness of person-fit statistics to detect social desirability or exaggerated protocols is unclear. Zickar and Drasgow (1996) found that a person-fit approach identified a higher number of faking persons than when using a social desirability scale, whereas Ferrando and Chico (2001) showed that the validity scales (i.e., Lie scale and the Social Desirability scale) outperformed a person-fit statistic.

In person-fit research the incongruity between a person’s estimated trait level and his or her pattern of item responses is investigated (e.g., a high scoring examinee missing an easy item). However, as Reise and Flannery (1996) noted “most social desirable responding ... is either subtle (e.g., changing a Likert rating of 3 to 4 but not to 8) or consistent throughout the majority of the protocol (e.g., elevating all ratings by a constant). Such responding would seldom produce discrepancies large enough to detect” (p. 12).
As an alternative, methods may be used that are based on counting the strings of identical scores. These methods may be much more powerful to detect this type of aberrant response behavior. IRT-based statistics have been proposed that are sensitive to strings of consecutive similar scores, such as strings of all 1 scores or all 4 scores (Meijer & Sijtsma, 2001). These statistics may indicate misfit as a result of faking or unmotivated response behavior.

6.3.1 Scale Properties

Several authors have discussed that items should be as discriminating as possible (i.e., high item-test correlations) to increase the detection rate of person-fit indices (e.g., Meijer, Molenaar, & Sijtsma, 1994). However, the selection of items with high discrimination will result in item sets that are very homogeneous in content, that is, items will be selected in which a similar question is repeated in slightly different way (“Often down in the dumps” and “Often feels unhappy”). This results in a trait variable that is highly reliable but extremely narrow in content (Egberink & Meijer, in press; Reise & Flannery, 1996).

Another challenge is that to identify person-misfit one needs tests with items that are dispersed across the latent trait range and tests that are relatively long (e.g., Ferrando & Chico, 2001). Although, long personality batteries exist, there is a trend in clinical and health psychology, medicine, and psychiatry for making decisions about patients using short questionnaires containing at most 15 items (Fliege, Becker, Walter, Björner, Klapp, & Rose, 2005; Reise & Waller, 2009; Walter, Becker, Björner, Fliege, Klapp, & Rose, 2007). These short tests might result in low power for several person-fit statistics.

6.3.2 Emphasis on Simulated Data

In many person-fit studies the power of the different statistics has been investigated using simulated data. It is, however, unclear whether the results obtained in these studies generalize to different types of empirical data. For example, Emons (2008) simulated item score patterns that mimic careless response behavior and item score pattern that mimic the tendency to choose extreme response options for simulated tests consisting of 12 and 24 items. For a test consisting of 12 items (comparable to the length of our scales, see below), he found higher detection rates for carelessness than for extreme response behavior. Furthermore, Emons (2008) found comparable detection rates for tests with low and high discriminating power for carelessness (Emons, 2008, Table 2, for 12 items and careless response behavior on all items) but for extreme response behavior he found higher detection rates for lower
discriminating items than for higher discriminating items. This latter observation was explained through the observation that “the higher the item discrimination, the higher the probability of responses in the lowest category of difficult items, and likewise for the highest category of easy items” (p. 238). As a result, increasing the discriminating power of an item will result in smaller differences of expected scores under the null model and the observed extreme response pattern. This is an interesting observation and would imply that for personality questionnaires that measure a relatively broad construct person-fit statistics may be useful.

The results by Emons (2008) were obtained through simulated data and the critical values were determined in a simulated dataset of model-fitting response patterns. In psychological practice and especially in a personnel selection context, it is less straightforward to decide when a pattern should be labeled as unexpected. To be able to classify an item response pattern as aberrant, one would like to have data containing no aberrant response patterns, that is, data consisting of persons that filled out the questionnaire without presenting themselves in a favorable daylight. On the basis of this data critical values can be determined. However, in practice we often have data where many persons have the tendency to choose extreme responses. This may drastically affect the power results reported in Emons (2008).

6.4 Empirical Evidence of the Usefulness of Validity Indices

Most of the conducted studies on response distortion and faking used student samples and induced faking to assess whether persons can fake, to assess the frequency of faking, and to assess the impact of faking on the psychometric properties of a test (e.g., Ferrando & Chico, 2001; Pauls & Crost, 2004). There are only a few, but promising, studies using real data that applied both validity scales and person-fit indices and compared their relative usefulness or tried to explain inconsistent response behavior through psychological reasons with data from knowledgeable others and/or data from other sources.

In one of these studies, Woods, Oltmanns, and Turkheimer (2008) investigated several covariates that may explain misfitting response behavior on five subscales of the Schedule for Nonadaptive and Adaptive personality (SNAP; Clark, 1993). Misfitting response behavior was measured through the use of three SNAP validity scales (Rare Virtues, Deviance, and Variable Response Inconsistency) and the use of a person-fit statistic. Furthermore, they used information from peers with respect to the extent to which each person exhibited features of obsessive compulsive
personality disorder (implying orderliness and perfection) or borderline personality disorder (impulsive and emotionally erratic and having unstable images of themselves and others). For five subscales of the SNAP they found misfitting response behavior. This behavior was partly related to severe pathology, although they suggested that carelessness, haphazard responding, or uncooperativeness may have been other potential causes of aberrant behavior.

In another study aimed at the psychological causes of invalid testing protocols, Meijer, Egberink, Emons, and Sijtsma (2008) conducted an extensive study of children who responded to a popular measure of self-concept. They identified several children with inconsistent response behavior after repeated assessment. Additional information was obtained through observations during test administration and through interviews with their teachers to identify the causes of invalid test results. Meijer et al. (2008) concluded that for some children in the sample, item scores did not adequately reflect their trait level. Based on teachers’ interviews, this was found to be due most likely to a less developed self-concept and/or problems understanding the meaning of the questions. Thus, Meijer et al. (2008) showed that individuals who display poor person fit are not necessarily merely generating random or other faulty responses. Indeed, there may be interesting psychological reasons why an individual’s response pattern is inconsistent with a model.

Finally, Conrad, Bezruczko, Chan, Riley, Diamond, and Dennis (2010) used person-fit statistics to identify atypical forms of suicide risk. They identified a group of persons with suicide symptoms but with low scores on symptoms of internalizing disorders, which is a major risk factor for suicide. As a result of low scores on internalizing disorders this group may be not be identified when patterns are not checked for aberrant responses.

In conclusion, there is some evidence that different types of validity indices may be useful to detect invalid test behavior, but for a practitioner it is unclear how different response distortion detection techniques are related to each other when applied to empirical data and how powerful different techniques are. Although one may argue that person-fit statistics may not be sensitive to faking or social desirable responding (impression management), it is an empirical question whether this is indeed the case. Impression management may only affect the answers to those items that are formulated in a socially desirable way, and may not affect the answers to the more neutral formulated items. To obtain more insight into the usefulness of different validity indicators in practice, we conducted an empirical study. More specifically, the aim of this study was to compare the performance of different
validity indicators using empirical data obtained in a personnel development and personnel selection context.

6.5 Method

6.5.1 Instruments

 Connector Big Five Personality Questionnaire (ConnP)
The ConnP (Schakel, Smid & Jaganjac, 2007) is a computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. The questionnaire is used in both a selection context and a career development context as a global assessment of the Big Five factors. It consists of 72 items, distributed over five scales (Emotional Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a five point Likert scale. The answer most indicative for the trait being measured is scored “5” and the answer least indicative for the trait is scored “1”. For this study we selected the Emotional Stability (15 items), Extraversion (15 items), Openness (12 items), and Conscientiousness scale (15 items). The ConnP is based on a Dutch version of the Workplace Big Five Profile constructed by Howard and Howard (2001), which is based on the NEO-PI-R and adapted to workplace situations. For the Dutch version, both conceptual analyses and exploratory factor analyses showed the Big Five structure (Schakel, et al., 2007). Also, recent research showed that the psychometric quality of the scales was acceptable (Egberink, et al., 2010).

6.5.2 Participants and Procedure
Data were collected between September 2009 and March 2010 in cooperation with a Dutch human resources assessment firm whenever a personality measure was administered to a client. We distinguished two groups: (1) the personnel selection group (applicants who apply for a job at an organization), and (2) the career development group (persons already working for the organization, and completing the ConnP as part of their own personal career development). Because the interests are higher for persons in the personnel selection group compared to the development group, we expect that persons in the personnel selection group have a higher tendency to respond in a social desirable way than in the development group. Because the interests are lower for the career development group, we expect that these persons will have a higher probability of unexpected responses due to random
mistakes, for example resulting from concentration loss or motivation loss, than persons in the selection group.

The personnel selection group consisted of 2217 persons ($M_{\text{age}} = 31.1$, $SD = 8.65$); 66.3% men and most persons were White. 50.0% of the participants had a university degree, 28.2% had higher education, and 20.8% secondary education; for 1.0% educational level was unknown. The career development group consisted of 1488 persons ($M_{\text{age}} = 39.8$, $SD = 9.28$); 54.4% men and most persons were White. 25.7% of the participants had a university degree, 51.4% had higher education, and 21.4% secondary education; for 1.5% educational level was unknown. Because the two groups differed with respect to the variables gender, educational level, and age, we checked whether systematic differences in mean scores on the personality scales still exist after keeping the variables gender, educational level, and age constant\(^\text{1}\). This was the case, which indicates that the differences in mean scores are likely caused by the employment status of the participants. These results correspond with findings in the literature that there are differences in response behavior between applicants and incumbents (e.g., Robie, Zickar, & Schmit, 2001; Weekley, Ployhart, & Harold, 2004).

### 6.5.3 Analyses

#### Person Fit

Given that the items in a questionnaire are ordered from most popular to least popular, a simple and powerful person-fit statistic is the number of Guttman errors. For dichotomous items, the number of Guttman errors equals the number of 0 scores preceding a 1 score in a score pattern. Thus, the pattern (10110010) contains five Guttman errors. A drawback of this statistic, however, is that it is confounded with the total score (Meijer, 1994). For polychotomous items Emons (2008), therefore, proposed a normed version of the number of Guttman errors:

$$G_N^p = \frac{G^p}{\max\{G^p | X_+\}}$$

\(^\text{1}\) In a 2 (male, female) x 3 (university, higher education, secondary education) x 3 (younger than 30 years of age, 31 through 38 years of age, older than 39 years of age) design, 18 independent samples $t$-tests were conducted. Fisher’s combined probability test (e.g., Littell & Folks, 1971) was used to combine the $p$-values of these 18 independent tests to obtain an overall $p$-value to test whether systematic differences in mean scores between two groups still exist when the background variables were kept constant. The overall $p$-value for each scale was smaller than .0005.
In this statistic the number of Guttman errors ($G^p$) is weighted by its maximum value given the sum score (for details see Emons, 2008). $G^p_N$ ranges from 0 (i.e., no misfit) through 1 (i.e., maximum misfit); for perfect response patterns (i.e., all 1 scores or all 5 scores) the statistic is undefined. Although, Emons (2008) found in a simulation study that the power of the simple number of Guttman errors was higher than the normed number of Guttman errors, we prefer using the normed version because it is less confounded with the total score than the simple number of Guttman errors.

**Social Desirability Scale**

Within a large number of organizations, besides filling out the ConnP items, persons also filled out ten ‘self-image’ items. These items are used to assess a person’s tendency to respond in a social desirable way. The items are included in the ConnP and are administered together with the content scale items. The self-image items are scored on a five point Likert scale where the answer most indicative for social desirability is scored “5” and the answer least indicative is scored “1”. An example of an item is “Is not driven to climb higher up on the career ladder” (reversed scored). A T-score of 55 (comparable with a total score of 37 or higher) indicates a high level of social desirable answering and a T-score of 65 or higher (comparable with a total score of 42 or higher) indicates extreme social desirable answering. The social desirability scale was filled out by 1570 persons from the personnel selection group, and by 363 persons from the career development group.

**Inconsistency Index**

Following the reasoning of the VRIN scale of the MMPI-2-RF (Ben-Porath & Tellegen, 2008), we constructed an inconsistency index based on the ConnP items. Although the ConnP is used for assessment at the factor level, each factor is divided into facets consisting of three items. For each factor, we selected facets that consist of three items that are semantically similar. From the Emotional stability scale we selected the facet Intensity that determines how easily we get angry and the facet Rebound time that determines how much time we need to rebound from setbacks; from the Extraversion scale we selected the facet Taking charge that measures the degree to which we assume a leadership role and the facet Directness that determines the degree to which we express our opinions directly; from the Conscientiousness scale we selected the facet Perfectionism that measures striving for perfect results and the facet Concentration that measures concentrating on a particular task; and from the Openness scale we selected the facet Imagination that
determines the number of new ideas and applications we think up. For the inconsistency index, we calculated for each of these seven facets the absolute differences between the responses on the three items (i.e., the difference between the item responses on item 1 and item 2, item 1 and 3, and item 2 and 3). The sum of these differences was taken as a measure for inconsistent response behavior. The higher the score, the more inconsistent the person responded. The minimum score equaled 0 and the maximum score equaled 56. An example of three semantically similar items from Extraversion scale is “Is reserved in expressing his/her opinion” (reversed scored), “Immediately says what he/she thinks of something”, and “Keeps his/her criticism to him/herself” (reversed scored).

Proportion of 5 Scores
To assess whether persons have a tendency to respond in the extreme response category, we calculated a simple statistic, namely the proportion of 5 scores. The higher the proportion, the more a person responded in the extreme response category. A large number of 5 scores may also indicate a valid score for persons with a high trait score. However, we hypothesize that a large number of 5 scores across different scales may indicate social desirability, faking, and/or extreme response behavior.

Power of the Validity Indicators
To obtain an impression of the power of the different methods we added 10% simulated response patterns to the data. We simulated two kinds of response distortion: (1) random response behavior where the realization of each score (i.e., 1, 2, 3, 4, or 5) had an equal probability of \( p = .2 \), and (2) extreme response behavior or social desirable responding where each respondent had an equal probability of \( p = .5 \) to obtain a 4 score or a 5 score. We simulated these data for two content scales, the Conscientiousness scale and the Emotional Stability scale, and for the Social Desirability scale. We added the simulated data to the original data, calculated the different statistics and determined the proportion of simulated patterns that belonged to the 10% most aberrant score patterns. This procedure was replicated 10 times.

---

2 Because of copyright restrictions we can only provide one example of a used facet to construct the inconsistency index.
6.6 Results

6.6.1 Descriptive Statistics Personality Scales

Scale means, standard deviations, and Cronbach’s α for the Emotional Stability, Extraversion, Openness, and Conscientiousness scale for the two groups are reported in Table 6.1. Cronbach’s α is similar for each scale across the groups. In general, the differences in mean score are large between the personnel selection group and the career development group, resulting in a large effect size (i.e., \( d = .85 \)) for the Conscientiousness scale, a medium to large effect size (i.e., \( d = .68 \)) for the Emotional Stability scale and a small to medium effect size (i.e., \( d = .36 \)) for the Extraversion scale. These results are in line with the literature. In their meta-analytic study, Birkeland, Manson, Kisamore, Brannick and Smith (2006) reported effect sizes of \( d = .45 \) for Conscientiousness and \( d = .44 \) for Emotional Stability between the mean scores of applicants and non-applicants.

<table>
<thead>
<tr>
<th>Scale characteristic</th>
<th>α</th>
<th>M</th>
<th>SD</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Stability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selection</td>
<td>.81</td>
<td>60.28</td>
<td>6.47</td>
<td>-.36</td>
<td>.01</td>
</tr>
<tr>
<td>career</td>
<td>.85</td>
<td>56.75</td>
<td>8.06</td>
<td>-.45</td>
<td>.10</td>
</tr>
<tr>
<td>Extraversion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selection</td>
<td>.81</td>
<td>60.27</td>
<td>6.53</td>
<td>-.53</td>
<td>.45</td>
</tr>
<tr>
<td>career</td>
<td>.80</td>
<td>58.50</td>
<td>7.27</td>
<td>-.41</td>
<td>-.04</td>
</tr>
<tr>
<td>Openness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selection</td>
<td>.82</td>
<td>48.07</td>
<td>5.48</td>
<td>-.26</td>
<td>-.07</td>
</tr>
<tr>
<td>career</td>
<td>.84</td>
<td>46.81</td>
<td>6.39</td>
<td>-.57</td>
<td>.80</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selection</td>
<td>.81</td>
<td>62.06</td>
<td>6.48</td>
<td>-.49</td>
<td>.20</td>
</tr>
<tr>
<td>career</td>
<td>.81</td>
<td>57.82</td>
<td>7.57</td>
<td>-.38</td>
<td>.12</td>
</tr>
</tbody>
</table>

Note. \( M \) = mean scale score; selection = personnel selection group; career = career development group. 
\( \alpha \) = coefficient alpha.

6.6.2 Descriptive Statistics Response Validity Indicators

Table 6.2 shows the mean scores on each response validity indicator for both groups. The career development group has a higher mean number of \( G^D_N \) for each
personality scale compared to the personnel selection group. One explanation may be that for the personnel selection group 77%-82% of the responses on the four scales was in category 4 and 5, compared to 68%-73% for the career development group. Thus, persons in the personnel selection group chose more often extreme answers. Because person-fit statistics, like $G^p_N$, are less suited to detect extreme response patterns, the mean number of $G^p_N$ will also be lower in the personnel selection group. Also note that the mean score on the social desirability scale was higher for the personnel selection group compared to the career development group (Cohen’s $d = .72$) and that the mean score on the inconsistency index was higher for the career development group compared to the personnel selection group (Cohen’s $d = .55$). This supports the idea that the more extreme response patterns were present in the personnel selection group and that the more inconsistent patterns were present in the career development group.

Table 6.2

| Mean scores on each response distortion detection technique for both groups. |
|-------------------------------|-------------------|-------------------|
|                               | personnel selection | career development |
|                               | $M$    | $SD$ | $M$    | $SD$ |
| $G^p_N$ _EMS                  | .108   | .08  | .121   | .08  |
| $G^p_N$ _EXT                  | .107   | .08  | .146   | .10  |
| $G^p_N$ _OPE                  | .096   | .08  | .113   | .10  |
| $G^p_N$ _CON                  | .116   | .09  | .129   | .09  |
| social desirability           | 38.84  | 4.10 | 36.30  | 5.70 |
| prop. #4s EMS                 | .43    | .19  | .40    | .19  |
| prop. #5s EMS                 | .34    | .24  | .28    | .23  |
| prop. #4s EXT                 | .46    | .19  | .43    | .19  |
| prop. #5s EXT                 | .32    | .22  | .30    | .22  |
| prop. #4s OPE                 | .47    | .21  | .45    | .22  |
| prop. #5s OPE                 | .30    | .25  | .27    | .25  |
| prop. #4s CON                 | .42    | .20  | .40    | .19  |
| prop. #5s CON                 | .40    | .25  | .31    | .23  |
| inconsistency                 | 15.66  | 5.42 | 17.82  | 5.68 |

Note. $G^p_N$ = normed Guttman errors; EMS = Emotional Stability; EXT = Extraversion; OPE = Openness; CON = Conscientiousness; prop. = proportion.
We also calculated the proportion of persons with a high score on both the personality scales and the social desirability scale. Scoring high on both the personality scales and the response distortion scale indicates response distortion. In the personnel selection group 45.8% of the persons with a high score on three personality scales also had a high score on the social desirability scale, in the career development group this was 55.6%. However, the proportion of persons with a high score on both the four personality scales and the social desirability scale equaled 84.2% for the personnel selection group and 33.3% for the career development. Thus, in the selection context a high score on the four subscales is strong evidence for response distortion.

6.6.3 Comparison of Validity Indicators

Figure 6.1 displays the number of response patterns that were classified as invalid in the personnel selection group (left panel) and in the career development group (right panel) using the different response validity indicators and combinations of these indicators. Based on the social desirability and inconsistency scores, a pattern was classified as invalid when it belonged to the 10% most extreme scores. Based on $G^p_N$ and the proportion of number of 5 scores for each of the four personality scales, a pattern was classified as invalid when the response patterns of a person belonged to the 10% most extreme scores for at least three out of four content scales.

![Figure 6.1: Venn diagram showing the overlap of the number of response patterns in the personnel selection group (left panel) and the career development group (right panel) classified as invalid using different validity indicators. Note. $G^p_N =$ normed Guttman errors with regard to the involved scale; SD = social desirability; inconsist = inconsistency index; prop #5s = proportion of numbers of 5 scores.](image)
scales. Figure 6.1 shows that there is not much overlap between the classifications of invalid score patterns using the different statistics. Thus, the different methods detect different types of invalid response patterns. The largest overlap was between \( G^p_N \) and the inconsistency index, and the social desirability scale and the proportion of #5 scores.

### 6.6.4 Power of the Different Validity Indicators in Real Empirical Data

Table 6.3 displays the average detection rates of the different validity indicators for the 10 random response datasets and the 10 extreme response datasets for the Conscientiousness scale and the Emotional Stability scale for both groups. The indicators that are sensitive to inconsistent response behavior (i.e., \( G^p_N \) and the inconsistency index) have the highest detection rate in the random response data and the statistics that are sensitive to extreme response behavior or social desirable responding (i.e., the social desirability score and the proportion of 5 scores) have the highest detection rates for the extreme response data.

<table>
<thead>
<tr>
<th>Table 6.3</th>
<th>Mean detection rates of the different validity indicators for 10 random response datasets and 10 extreme response datasets for Conscientiousness and Emotional Stability for both groups.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>external application group</td>
</tr>
<tr>
<td></td>
<td>( G^p_N )</td>
</tr>
<tr>
<td>random response data (CON)</td>
<td>.63 (.03)</td>
</tr>
<tr>
<td>extreme response data (CON)</td>
<td>.02 (.01)</td>
</tr>
<tr>
<td>random response data (EMS)</td>
<td>.68 (.01)</td>
</tr>
<tr>
<td>extreme response data (EMS)</td>
<td>.05 (.02)</td>
</tr>
</tbody>
</table>

*Note.* \( G^p_N \) = normed Guttman errors with regard to the involved scale; SD = social desirability; incons = inconsistency index; #5s = proportion of numbers of 5 scores; EXT = Extraversion; EMS = Emotional Stability; standard deviations are displayed between brackets.
6.7 Discussion

Although a particular questionnaire can be a good measure of a psychological construct for a group of persons, it may be a poor measure of the construct for a particular individual (Ben-Porath & Waller, 1992). Therefore, information about the consistency of answering behavior on questionnaires can be of great help to practitioners in different fields of psychology. How should we use validity indicators in practice?

Our results showed that different validity indicators detect different types of distorted response patterns. The social desirability scale score, the proportion of number 5 scores, and the combination of high scale scores and a high score on the social desirability scale can be of help to detect exaggerated responses. Person-fit statistics like \( G^p_N \) are especially useful to detect unlikely score patterns. Furthermore, practitioners should realize that the power of person-fit statistics depend on the specific characteristics of the empirical data analyzed. When, such as in our case, there are many extreme scores, detection of faking through person-fit indices is useless.

6.7.1 How Can We Use Person-fit Statistics in Practice?

Quality measures like reliability and validity indices are defined at the group level and do not provide much information about the quality of individual measurement. Although different types of validity scales are often used in (especially clinical) practice to assess the quality of individual test scores, IRT-based person-fit statistics are seldom used in practice. We think that information from inconsistent response patterns combined with other (statistical) information can be of great help to identify suspicious test scores of individuals and groups of individuals to check measurement quality.

Routinely calculating person-fit statistics and evaluating them at the group level can be useful to check for scoring and other administration errors. For example, to check for hand-scoring errors by psychologists Simons, Goddard, and Patton (2002) investigated hand-scoring errors in aptitude, clinical, and personality measures and identified serious error rates for both psychologist and client scorers across all tests investigated. Also the use of computer-based tests and the automatic scoring of these tests may lead to serious errors as a result of misgrading. In the educational testing field this has led to serious problems and legal procedures (see Strauss, 2010, for an overview). Checking the likelihood of item score patterns can alert
practitioners and testing companies to these kinds of scoring and administration errors.

On an individual level, checking for invalid protocols results in immediate feedback. In clinical practice validity scales are often used, but also using person-fit statistics may reveal interesting idiosyncrasies (e.g., Conrad et al., 2010). In the context of computer-based diagnostic testing, calculating and reporting person-fit statistics is very easy. Testing may even be adapted and lengthened when it is clear that a patient produces very inconsistent response patterns. However, as we showed in our empirical analysis, person-fit statistics like the number of Guttman errors are sensitive to particular types of idiosyncratic answering behavior and the use of different sources of statistical and substantive information is necessary before we can conclude that individual test scores can not be trusted.
References


References


References


References


Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*, 125-146.


Summary

Applications of Item Response Theory to Non-Cognitive Data

The overarching aim of this thesis is to illustrate the usefulness of item response theory (IRT) for practical purposes at the individual person level. Psychological measurement instruments are used to make important decisions (e.g., school admission or personnel selection). In such cases, high-quality measurement instruments are required. IRT can be of help to construct and to evaluate the psychometric quality of psychological tests and can also help researchers to improve measurement at the individual person level. Although IRT originally has been used to determine the quality of cognitive tests (e.g., aptitude testing), in recent years, the use of IRT has become more popular in the non-cognitive domain (e.g., personality).

After a general introduction to IRT (Chapter 1), Chapter 2 illustrates the application of both parametric and nonparametric IRT models to evaluate the psychometric properties of a non-cognitive self-report questionnaire, the Self-Perception Profile for Children (SPPC; Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004). This questionnaire has been constructed according to classical test theory. Results showed that most subscales (consisting of 6 items) formed weak scales and that measurement precision was relatively low, although measurement precision differed across latent trait values within a particular subscale and also across scales. Some SPPC constructs can be measured by only 4 out of 6 items, whereas other constructs need more than 6 items. Furthermore, results showed that strong SPPC scales consisted mostly of items that repeat the same item content, and as a result these scales measured a relatively small construct. In Chapter 2, it is concluded that researchers should be careful when evaluating the quality of non-cognitive measures on the basis of their relation with an underlying latent trait, and that researchers need to find a balance between the bandwidth of the construct of interest and the measurement precision.
In the Chapters 3 and 4, the usefulness of IRT-based person-fit methodology in the non-cognitive domain is illustrated. Using person-fit statistics it is possible to detect item score patterns that do not fit an IRT model. Although many person-fit statistics exist, there are few studies that illustrate the usefulness of these statistics in practice. In Chapter 3, the same data were used as in Chapter 2. Results showed that for some children scale scores should be interpreted carefully. Combined information from person-fit statistics, observation, interviews, and self-concept theory showed that similar score profiles may have a different interpretation. Misfit was most likely due to a less developed self-concept and/or problems understanding the meaning of the questions. In Chapter 4, the inconsistency of individual items score patterns in a computerized adaptive test for personality is illustrated using the cumulative sum procedure (CUSUM; van Krimpen-Stoop & Meijer, 2002). In this chapter, it is also shown that similar estimated trait values may have different interpretations. Persons with an inconsistent item score pattern responded more often in the extreme response categories (i.e., totally disagree and totally agree). Combined information from the CUSUM, other personality measures, and interviews suggested that these inconsistent response patterns were the result of a tendency to black-and-white thinking.

In Chapter 5, a mixture IRT model (e.g., Rost, 1990; Rost & Langeheine, 1997) is applied to a Conscientiousness scale in a career development context to assess differences in response scale usage at the group level and to compare the predictive validity of the mixture IRT trait estimates with unidimensional IRT trait estimates. Four different groups were identified based on their response behavior; they mainly differed with respect to their scores on the subscales Perfectionism and Concentration. Furthermore, results showed that Conscientiousness may be qualitatively different for different groups of persons and that the predictive validity of the test scores improved for persons in different classes as suggested by using the mixture IRT model as compared to fitting a unidimensional IRT model.

In Chapter 6, the usefulness of different types of validity indicators (i.e., a social desirability scale, an inconsistency index, a person-fit statistic, and an extreme response index) to detect different types of aberrant response behavior is investigated in an external application group, an internal application group, and a career development group. The results showed that the external and internal application groups contained more often extreme response behavior, whereas the career development group contained more often random responding. Furthermore, there was not much overlap between the detection of aberrant response behavior by the different validity indicators. This implies that different approaches are sensitive
to different types of invalid response behavior. Furthermore, the results showed that the person-fit statistic and the inconsistency index are useful to detect random response behavior and that the social desirability scale and the extreme response index (i.e., proportion extreme responses) are useful to detect extreme response behavior.

The chapters in this thesis illustrate that IRT can be of help in the non-cognitive domain to evaluate the psychometric properties of a questionnaire, and to improve measurement and diagnostics at the individual level. Various statistical methods combined with auxiliary information from other inventories, observation, and interviews can be helpful to interpret the validity of test scores.
Samenvatting (Summary in Dutch)
Toepassingen van Item Respons Theorie op Niet-Cognitieve Data

In dit proefschrift wordt het gebruik van item respons theorie (IRT) om vragenlijsten in het niet-cognitieve domein te analyseren onderzocht. Het hoofddoel van dit proefschrift is om het nut van IRT voor praktische doeleinden te illustreren op het niveau van de respondent. Steeds vaker worden psychologische meetinstrumenten gebruikt om belangrijke beslissingen te nemen, denk hierbij aan de invloed van de CITO score op de schoolkeuze van kinderen of het gebruik van intelligentietests bij de selectie van personeel. Bij het gebruik van testscores voor dit soort beslissingen, moet men er dus op kunnen vertrouwen dat de scores goed gemeten zijn. Hoge kwaliteitseisen aan het meetinstrument zijn dan noodzakelijk. Met behulp van IRT kan aan deze eisen worden voldaan. IRT is van oudsher veel toegepast en wordt nog steeds veel toegepast in het cognitieve domein (bijvoorbeeld intelligentietests), maar IRT wordt de laatste jaren ook steeds vaker toegepast in het niet-cognitieve domein (bijvoorbeeld persoonlijkheidsmeting).

Na een algemene inleiding over IRT (Hoofdstuk 1), wordt in Hoofdstuk 2 het gebruik van zowel parametrische als niet-parametrische IRT geïllustreerd aan de hand van een analyse van de Competentie BelevingsSchaal voor Kinderen (CBSK; Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004). Deze vragenlijst is ontwikkeld op basis van de klassieke test theorie. De resultaten laten zien dat de meeste subschalen (bestaande uit 6 items) zwakke schalen zijn en dat de meetprecisie relatief laag is. Tevens verschilt de meetprecisie voor verschillende latentere trek waarden binnen een schaal en per schaal. Sommige constructen kunnen nauwkeurig gemeten worden met slechts 4 van de 6 items, terwijl andere constructen meer dan 6 items nodig hebben. Verder wordt in dit hoofdstuk besproken dat men zich niet alleen moet richten op de meetprecisie bij het beoordelen van de kwaliteit van een instrument. Zo lieten de resultaten zien dat goede CBSK schalen voornamelijk bestaan uit semantisch gelijke items hetgeen resulteert in het meten van relatief smalle constructen. De conclusie van dit hoofdstuk is dan ook dat de subschaalscores van de CBSK zorgvuldig en voorzichtig moeten worden
geïnterpreteerd en dat onderzoekers een balans moeten vinden tussen de inhoud van de items en de meetprecisie van het instrument.

In Hoofdstuk 3 en 4 wordt de bruikbaarheid van op IRT gebaseerde “person-fit” methoden om inconsistente antwoordpatronen bij persoonlijkheids meting te detecteren geïllustreerd. Er bestaan verschillende “person-fit” methoden, maar er zijn weinig studies die de bruikbaarheid van deze methoden in de praktijk illustreren. In Hoofdstuk 3 worden dezelfde CBSK data gebruikt als in Hoofdstuk 2. De resultaten laten zien dat voor sommige kinderen de schaalscores voorzichtig en zorgvuldig moeten worden geïnterpreteerd. Informatie verkregen door “person-fit” methoden gecombineerd met informatie verkregen door observatie, interviews en theorieën over zelfbeeld laten zien dat gelijke scoreprofielen een verschillende interpretatie kunnen hebben. Gebaseerd op interviews met de leerkrachten bleek dat dit te maken had met een minder ontwikkeld zelfbeeld en/of problemen met het begrijpen van de vragen. In Hoofdstuk 4 wordt de inconsistentie van individuele item score patronen op een computeradaptieve vragenlijst voor persoonlijkheid onderzocht. Hiervoor is gebruik gemaakt van de “cumulative sum” procedure (CUSUM; van Krimpen-Stoop & Meijer, 2002). Ook in dit hoofdstuk wordt duidelijk dat gelijke testscores een verschillende betekenis kunnen hebben. Mensen met een inconsistent antwoordpatroon bleken vaak te antwoorden in de extreme antwoordcategorieën (helemaal mee oneens en helemaal mee eens). Gecombineerde informatie van de CUSUM, andere persoonlijkheids vragenlijsten en interviews suggereert dat deze inconsistentie antwoord patronen het resultaat zijn van een neiging om de wereld te zien in termen van extremen en zwart-wit denken.

In Hoofdstuk 5 wordt het gebruik van de antwoordcategorieën op groeps niveau in kaart gebracht. Dit wordt gedaan door een mixture IRT model toe te passen op een Consciëntieusheidschaal in een loopbaanontwikkelingscontext. In dit hoofdstuk wordt zowel gekeken naar de verschillen in het gebruik van de antwoordcategorieën als naar de predictieve validiteit van scores geschat met een mixture IRT model en geschat met een eendimensioneel IRT model. Er worden vier groepen onderscheiden op basis van hun antwoordgedrag en deze groepen verschillen met name in het gebruik van de antwoordcategorieën met betrekking tot de schalen Concentratie en Perfectionisme. De resultaten laten zien dat Consciëntieusheid kwalitatief verschillend kan zijn voor verschillende groepen. Deze verschillen in antwoordgedrag leiden ertoe dat de predictieve validiteit van testscores van personen in de verschillende groepen, zoals gevonden door middel van een mixture IRT model, verbetert ten opzichte van de predictieve validiteit van testscores geschat met een eendimensioneel IRT model.
In Hoofdstuk 6 wordt de bruikbaarheid van verschillende type validiteitsindicatoren (sociaal wenselijkheidsschaal, inconsistentie index, “person-fit” methode en index voor extreem antwoordgedrag) om verschillende typen afwijkend antwoordgedrag te detecteren onderzocht in een externe sollicitatiegroep, een interne sollicitatiegroep en een loopbaanontwikkelingsgroep. De resultaten laten zien dat de externe en interne sollicitatiegroep vaker extreem antwoordgedrag vertonen, terwijl de loopbaanontwikkelingsgroep vaker willekeurig antwoordt. Verder is er weinig overeenkomst in de detectie van afwijkend antwoordgedrag door verschillende validiteitsindicatoren. Dit betekent dat de verschillende methoden verschillende typen afwijkend antwoordgedrag detecteren. De resultaten laten zien dat willekeurig antwoordgedrag het beste gedetecteerd kan worden met de “person-fit” methode en de inconsistentie index. Extreem antwoordgedrag of sociaal wenselijk antwoorden kan het beste gedetecteerd worden met de sociaal wenselijkheidsschaal en de index voor extreem antwoordgedrag (proportie antwoorden in de hoogste antwoordcategorie).

De studies in dit proefschrift leiden tot de conclusie dat IRT behulpzaam kan zijn bij het in kaart brengen van de psychometrische eigenschappen en bij het verbeteren van de individuele diagnostiek van vragenlijsten in het niet-cognitieve domein. Testscores kunnen op verschillende manieren tot stand komen. Verschillende statistische methoden in combinatie met aanvullende informatie uit andere vragenlijsten, observatie en interviews kunnen helpen om de validiteit van die testscores te interpreteren. Daarnaast zijn verschillende validiteitsindicatoren geschikt om verschillende typen afwijkend antwoordgedrag te detecteren.
Dankwoord (Acknowledgements)

In 2006 besloten PiCompany en de Universiteit Twente een samenwerkingsverband aan te gaan, met als belangrijkste element het financieren en faciliteren van een AiO project voor de duur van vier jaar. Ik was de gelukkige kandidaat die in november 2006 met dit project aan de slag mocht gaan. Inmiddels zijn de vier jaar voorbij gevlogen, waarin ik veel geleerd heb en hard gewerkt heb aan de mooie uitdaging om wetenschap en praktijk dichter bij elkaar te brengen. Het resultaat hiervan ligt voor u, namelijk mijn proefschrift. Hierin is te lezen hoe statistische modellen en methoden een bijdrage kunnen leveren aan en gebruikt kunnen worden in de (psychologische) praktijk. Dit resultaat heb ik niet alleen bereikt, het is tot stand gekomen door de ondersteuning van en de samenwerking met verschillende mensen en organisaties. Hen wil ik daarvoor graag bedanken.

Allereerst wil ik mijn promotor Rob Meijer bedanken, die (na mij) ongetwijfeld het meest heeft bijgedragen aan dit proefschrift. Onze gedeelde interesse voor de psychologie, het psychologisch testen en de psychometrie was de basis voor een goede en resultaatgerichte samenwerking. Ik heb met heel veel plezier met je samengewerkt en veel van je geleerd, met name hoe de psychometrie een zinvolle bijdrage kan leveren aan de praktijk en hoe statistische resultaten zinvol geïnterpreteerd kunnen worden vanuit de psychologische invalshoek. Naast het werken aan de artikelen, betrok je me ook bij het onderwijs. Af en toe college ‘Testtheorie & Psychodiagnostiek’ geven, helpen bij practica en samen masterstudenten begeleiden tijdens hun afstuderen, vond ik geweldig en waren voor mij leerzame ervaringen. Dankjewel dat je me die mogelijkheid hebt geboden. Jouw vertrek naar Groningen, halverwege mijn project, veranderde gelukkig weinig aan onze samenwerking, mede door jouw goede bereikbaarheid en snelle reactie via de mail. In plaats van een deurtje verder voor overleg, ging ik nu een dag naar Groningen. Rob, en ook Ilse, bedankt voor de gezellige momenten daar in het ‘hoge noorden’.

Als tweede wil ik mijn copromotor Bernard Veldkamp bedanken. Bernard, waar ik door Rob vanuit de psychologie en de psychometrie ondersteund werd, steunde jij me vooral bij de technische aspecten van mijn onderzoek. Ik heb veel geleerd van jouw kennis over computer adaptief toetsen. Ook jij bood me de mogelijkheid om te
participeren in het onderwijs door colleges te geven en samen afstudeerders te begeleiden. Dankjewel daarvoor. Door het vertrek van Rob werd je mijn dagelijks begeleider in Enschede en intensieveer ons contact. Dit heb ik als prettig ervaren; altijd was je bereid om met een kritische blik naar mijn werk te kijken en om mijn klankbord te zijn, zowel werkgerelateerd als niet-werkgerelateerd.

Natuurlijk was dit alles niet gelukt zonder PiCompany, zij hebben immers het gehele project gefinancierd. In het bijzonder wil ik Martijn van der Woude, Nico Smid en Lolle Schakel bedanken voor hun betrokkenheid bij het project. Dankzij jullie kon ik een kijkje nemen in de keuken van de praktijk en had ik de beschikbaarheid over grote sets “echte” data, daar heb ik veel van geleerd en ben ik jullie dankbaar voor. Ik kijk met heel veel plezier terug op onze samenwerking.

Verder bedank ik graag de leden van de beoordelingscommissie, Marise Born, Klaas Sijtsma en Hans Hoekstra, voor het aandachtig lezen van mijn proefschrift en hun aanwezigheid tijdens mijn promotie. Tevens bedank ik de overige leden van de promotiecommissie voor hun aanwezigheid.

Graag bedank ik alle (ex-)collega’s van de afdeling OMD voor hun belangstelling, het kritisch meedenken en de gezelligheid. Een aantal mensen wil ik in het bijzonder bedanken. Cees Glas, bedankt voor het delen van je kennis over de item respons theorie en computer adaptief toetsen, maar ook voor je gastvrijheid zodat ik na het vertrek van Rob in Enschede kon blijven werken. Josine en Marianna, bedankt dat jullie me als paranimfen willen bijstaan gedurende mijn verdediging. Lorette en Birgit, bedankt voor jullie handige tips en adviezen, het regelen van allerlei zaken en vooral bedankt voor de gezelligheid. Rinke, bedankt voor het genereren van de data die gebruikt zijn voor het tweede deel van Hoofdstuk 6.

Tevens wil ik graag de collega’s bij PiCompany bedanken; de PiCompanen van de afdelingen Development, ICT en Servicedesk wil ik bedanken voor hun hulp, steun, interesse en gezelligheid.

Buiten het werk om stonden mijn familie en vrienden ook altijd voor me klaar. Daarom bedank ik hier mijn familie en vrienden voor hun steun, interesse, gezelligheid en niet te vergeten voor hun geduld gedurende mijn promotietraject. Vaak was het ‘drukdruk’, maar een gezellig samenzijn en weer eens bijkletsen onder het genot van een kop thee was dan altijd een zeer aangename onderbreking, net als het gezellig samen fitnessen, een uurtje RPM en/of Bodysteps.

Tot slot bedank ik mijn ouders voor hun onvoorwaardelijke geloof in mijn kunnen, hun steun, belangstelling en liefde. Mam, bedankt dat ik jouw perfectionisme en doorzettingsvermogen heb; eigenschappen die je als promovenda heel goed kunt gebruiken. Pap, bedankt voor de fijne gesprekken en je bereidheid altijd mee te willen denken, onder andere hoe we je netwerk kunnen gebruiken voor het verzamelen van data, zoals we gedaan hebben bij het verzamelen van de data die gebruikt zijn in de Hoofdstukken 2, 3 en 4. Rosa, Job en Gijs, bedankt voor jullie belangstelling, gezelligheid en vooral ook voor jullie hulp bij het tekenen van de intelligentie items voor PiCompany. Huub, bedankt dat je van mijn kaft jouw kunstwerk hebt willen maken. Ik vond het erg gezellig om er samen mee bezig te zijn. Last but not least bedank ik mijn lieve vriend René voor zijn belangstelling, steun, begrip en liefde, ook voor alle avonden samen op de bank met de laptop op schoot.