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

I. J. L. Egberink (s0037826)

Begeleiders: dr. R. R. Meijer
dr. ir. B. P. Veldkamp

Enschede, augustus 2006

Universiteit Twente
Faculteit Gedragswetenschappen

DETECTION AND VALIDATION OF UNSCALABLE ITEM SCORE PATTERNS

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

Iris J. L. Egberink

University of Twente

Abstract

I illustrate the usefulness of person-fit methodology proposed in the context of item response theory in the field of personality assessment. First, I give a nontechnical introduction to existing person-fit statistics. Second, I analyze data from Harter's Self-perception Profile for Children (SPPC; Harter, 1985) in a sample consisting of children 8-12 years of age (N = 611) and argue that for some children the scale scores should be interpreted with care. Combined information from person-fit indices and from observation, interviews, and self-concept theory showed that similar score profiles have a different interpretation. For some children in the sample due to a less developed self-concept and/or problems understanding the wording of the questions, item scores did not adequately reflect their trait level. I recommend investigating the scalability of score patterns when using self-report inventories to withhold the researcher from wrong interpretations.

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

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, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). 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). 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 devoted to the psychometric characteristics of the statistics and tests (such as the correct sampling distribution) but there are very few studies that illustrate 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. Rudner, Bracey, and Skaggs (1996) remarked that "in general, we need more clinical oriented studies that find aberrant patterns of responses and then follow up with 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".

In the present study I try 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 study is organized as follows. First, I explain the usefulness of IRT to investigate the quality of individual item score patterns. Second, I provide a nontechnical background to person-fit analysis in the context of nonparametric IRT. Finally, I illustrate the practical usefulness of person-fit statistics in the context of personality assessment using the Self-perception Profile for Children (SPPC; Harter, 1985).

Item Response Theory and Individual Score Patterns

IRT Measurement Model

The use of IRT models in the personality domain is increasing. This is mainly due to the theoretical superiority of IRT to classical test theory (CTT). Although empirical similarities of tests and inventories constructed according to CTT and IRT do exist, IRT offers more elegant ways to investigate data than CTT. The detection of invalid test scores is an interesting example (Meijer, 2003). Reise and Henson (2003) provide 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 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 IRT models it is assumed that the probability of item

endorsement should increase as the trait levels increase, thus IRFs are monotonically increasing functions. In Figure 1 examples are given of several IRFs. More formally, the IRF, denoted $P_g(\theta)$, gives the probability of endorsing an item g (g = 1, . . ., k) as a function of θ. It is the probability of a positive response (i.e.,"agree" or "true") among persons with the latent trait value θ. For dichotomous items, $P_g(\theta)$ often is specified using the 1-, 2-, or 3-*parameter* logistic model (1-, 2-, 3PLM, see Embretson & Reise, 2000). These models are characterized by an s-shaped IRF. Examples are the IRFs 1, 2, and 3 in Figure 1.

Nonparametric IRT. In the present study, I use 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 1 all IRFs can be described by the MMH model, whereas the IRFs of the 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 the difference between parametric and nonparametric models). Another advantage is that the MMH is a relatively simple model that is easy to communicate with applied researchers.

The MMH model allows the ordering of persons with respect to θ using the unweighted sum of item scores (total score). Although many psychologists use the sum of the item scores or some transformation of it (e.g., *T* scores) without using any IRT model, they do not investigate and thus do not know if they can rank order persons according to their total score. Using the MMH, I first investigate if a model applies to the data before I use the total score to rank order

persons. Investigating the fit of the model also has the advantage that items can be identified that do not contribute to the rank ordering of persons.

The MMH is a probabilistic approach to the analysis of item scores which replaces the well-known Guttman (1950) model. According to the Guttman model it is not allowed that a subject endorses a less popular item while rejecting a more popular item. Obviously, this is an unrealistic requirement of response behavior. Probabilistic models such as the Mokken model allow deviations ("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, which I 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 a positively worded personality item having three ordered answer categories. It is assumed that the subject first ascertains whether he or she agrees enough with the statement to take the first item step. If not, the first item step equals 0, and the item score also equals zero. If the answer is affirmative, the item step equals 1, and the subject has to ascertain whether the second step 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 θ. Let x_g denote the polytomous score variable with realization h on item g and let $P_{gh}(\theta)$ denote the probability of a polytomous item score h on item g then the ISRF is defined as

$$P_{gh}(\theta) = P(x_g \geq h \mid \theta),\ g = 1, \ldots, k;\ \text{and}\ h = 0, \ldots, m.$$

It may be noted that $h = 0$ leads to a probability of 1 for each item, which is not informative about item functioning. This means that each item with $m + 1$ answer categories has m meaningful ISRFs. The MMH assumes that each of the ISRFs is monotone increasing in θ. Nondecreasingness of ISRF can be investigated by inspection of the observed item step score on the test score regression. Sometimes a rest score is used which is defined as the score on the other k-1 items without the score on the item g. The ISRF should be a monotonely nondecreasing function of the rest score. In Figure 2 examples of ISRFs 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 total (or rest) score for ordering respondents on θ.

To check whether the ISRFs are monotonically increasing several methods have been proposed. In this study I use the coefficient H_g for items ($g = 1, \ldots k$) and coefficient H for a set of items. Increasing values of H and H_g between .30 and 1.00 (maximum) mean that the evidence for monotone increasing ISRFs is more convincing, whereas values below .30 indicate violations of increasing ISRFs (for a discussion of these measures see for example, Meijer & Baneke, 2003 or Sijtsma & Molenaar, 2002). Furthermore, weak scalability is obtained if $.30 \leq H < .40$, medium scalability if $.40 \leq H < .50$ and strong scalability if $.50 \leq H < 1$.

Studying individual item score patterns.

When an IRT model gives a good description of the data it is possible to predict how persons at particular trait levels should behave when confronted with a particular set of test items. Let me illustrate this by means of Figure 3. For the sake of simplicity, I depicted five IRFs that do not intersect across the latent trait range. Assume that I have an estimate of someone's trait level to be $\theta = 0$ then the probability of endorsing item 1 equals .9 (most popular item) and

the probability of endorsing item 5 equals .1 (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 is thus unexpected and it may be questioned whether the total score of three has the same meaning for both patterns.

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 errors. 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 (1110101) consists of three Guttman errors. 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 then can be ordered according to decreasing popularity. A Guttman error consists of endorsing a so-called less popular item step while not endorsing a more popular item step. To illustrate this, consider a scale that consists of six items with four response alternatives (coded 1 through 4). This implies that there are three item steps per item (from 1-2, 2-3, and 3-4). Thus there are 6*3 = 18 item steps for each person. An example of a score pattern is (111111111011101101) which results in 10 Guttman errors.

Person-fit studies in the personality domain. Studies in a personality context investigated by means of simulated data whether person-fit statistics could be used as alternatives to social desirability and lying scales in order to identify dishonest respondents. Results were 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 "traitedness", Tellegen, 1988). They investigated the usefulness of a person-fit statistic to investigate traitedness 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. Care should be taken, however, in the interpretation of misfitting item score patterns. Reise and Waller (1993) discussed that interpreting misfitting item score patterns as an indicator of traitedness variation is difficult. Possible causes may be response faultiness, misreading, or random responding. Thus, many person-fit statistics do not allow the recovery of the mechanism that created the deviant item score patterns.

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 I combine information from person-fit statistics with auxiliary information from personality theory, and a respondent's personal history. Although in many studies it has been suggested that quantitative and qualitative information should be combined, there are few studies where this has been done (for an example, see Emons, 2003, chap. 6).

Method

Instrument

Data were analyzed from the official Dutch translation of Harter's (1985) Self-perception Profile for Children (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 on 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 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 four points scale. The answer most indicative for competence is scored 4 and the answer least indicative for competence gets a score 1.

To date, psychometric properties (multidimensional structure, invariance across groups) of the SPPC have been investigated mainly using CTT and factor analytical approaches. Veerman et al. (2004, p. 21- 25) showed a reasonable fit of the Dutch version of the SPPC of a five factor model with coefficient alpha for the subscales between .68 (BC) and .83 (PA).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 group.

Participants and procedure

Data were collected from 702 primary school children between 7 and 13 years of age. There were 391 mostly White girls and 311 mostly White boys (mean age = 9.82). These children attended primary schools in the east of the Netherlands. From this dataset I 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. There were five children older than 12 years of age, they were not removed from the data. The final sample consisted of 611 children, 343 girls and 268 boys (mean age 10.18). The research reported in this study was part of a larger project where routinely information was obtained from the children about their emotional and personal well-being.

Before the children completed the scales, standardized oral instructions were provided by the author. During the test administration, the author was available for further clarification. After test administration I analyzed the data (model fit), constructed a profile of the subtest scores for each child, and calculated the person fit statistics. The results were discussed with the teacher. During these conversations I explained the subtest profiles and also discussed possible explanations for inconsistent response behavior for those persons that were flagged as inconsistent by the person-fit statistic. Possible explanations are for example learning retardation, problems with reading comprehension skills, lexical processing speed or concentration problems.

Four of the five schools that participated in this study allowed me to re-administer the SPPC for children with inconsistent response behavior. This enabled me to determine the stability of the results. Before I re-administered the SPPC, I explained to the children that I randomly picked some children for research purposes. I did not explain that they produced irregular patterns because I would like to keep retest conditions as similar as possible as the first

time. The main difference was that at the second administration the children were tested in smaller groups than at the first administration.

Analysis

Model fit. I 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 calculate the fit of the item score patterns. I checked the assumptions of the MMH by inspecting the H and H_g coefficients and by inspecting ISRFs.

Because I suspected on the basis of the literature (Van den Bergh & van Ranst, 1998) and on the basis of my 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, I 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 file and investigate the H values across the different groups. This was done to ease the interpretation of misfitting item score patterns. When groups of persons are less scalable than others, individual misfit across groups may be more difficult to interpret. Furthermore, I investigated whether scale scores were similarly distributed across young and old children and 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 I restrict myself to descriptive and diagnostic information. Furthermore, I re-administered the SPPC for children with aberrant item score patterns. I 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 calculated for each subscale and then summed over subscales[1].

Re-administration of the SPPC. The SPPC was re-administered to children with Z_{GE} scores larger than 2.0. This value was based on the distribution of Z_{GE} scores resulting from the first administration (see Figure 4a) and eye-ball inspection of the item score patterns with Z_{GE} values at the right tail of the distribution by the author of this paper. Z_{GE} values smaller than 2.0 were less "strange" than Z_{GE} values larger than 2.0. This value also corresponds with the value I found when applying the rule of thumb for outliers of 1.5 times the difference between the first and the third quartile. On the basis of regression toward the mean I expect that, in general, Z_{GE} values will be smaller at the second administration than at the first administration. Therefore, I was especially interested in children that obtained approximately similar Z_{GE} scores at both test administrations.

Results

Descriptive statistics and scalability

Scale means, standard deviations, and internal consistency reliability (α) for each of the SPPC scales are reported in the Tables 1a and 1b. Mean scores for the boys were higher than for the girls on most scales, except for the BC scale and young children scored somewhat higher than old children except on the SA scale. Coefficient α was higher for girls than for boys and higher for old children than for young children. H and H_g values for the whole group are reported

in Table 2. From the table it can be concluded that with a few exceptions (in particular for the AC subscale) H_g and H values were larger than .3. Thus most items comply to the minimum requirements of the MMH model. Further inspection of the ISRFs showed no violations against increasing ISRFs.

In addition, I inspected the H and H_g values for boys and girls and young- and old children (not tabulated). Scalability was higher for the old children across all scales and higher for girls than for boys except for the BC scale (higher for boys) and the SC scale (equal). For old children I found for most scales medium scalability (H between .40 and .50), whereas for young children I found weak scales (H between .30 and .40). For girls I found mostly medium to strong scales with the exception of the AC and BC scales. For boys I 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 to use 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.

Person fit across different groups

Observation. A first inspection of the distribution of the Z_{GE} values (Figure 4a) showed that it was skewed to the right. Figures 4b and 4c present the distribution for young and old children. As expected, young children have on average higher Z_{GE} values, indicating more inconsistent behavior, than old children. I found a mean Z_{GE} = .265 for the young children and a mean Z_{GE} = -.211 for the old children, which is significant at a .001 level (t = 5.82; p=0.000). In

terms of effect sizes there is thus a difference in Z_{GE} scores of one half standard deviation.

Explanation. There may be a number of reasons for this finding. Responding to personality items which asks children to select statements that better describe them may be relatively complex for especially young children. They should understand the wording and should also have a similar frame of reference than old children. I observed that the wording of some items (in the Dutch translation I used) was problematic. During test administration young children asked questions about the wording of item 6 of the SA scale, item 3 of the BC scale, and item 3 of the GS scale. I found large differences between the H_g values for the items 4, 5, and 6 of the SA scale and the 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 for which the phrasing is too difficult, it may result in a random response and motivation 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). Although for very young children (6 and 7 years of age) age-appropriate assessments are used, where items are formulated more concretely and the bipolair formulation of the questions is eliminated (see e.g.,Van den Bergh & Rycke, 2003), my 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 between 13 and 20 years of age sometimes checked only 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 & Mac

Iver, 1989), whereas other researchers found an increasing differentiation of self-concept dimensions with age (e.g., Eccles, Wigfield, Harold, & Blumenfield, 1993; Eder, 1990; Marsh, Craven, & Debus, 1991, 1998). I 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".

Interesting in this respect is the shift that occurs in the proportions of endorsed response categories across age and gender. Table 3 shows that old children more often choose the categories 2 and 3 than young children. Moreover, old girls more often choose the 2 and 3 options than old boys. 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) give three explanations for children aged 7 and 8 for extreme response behavior: unrealistically high perceptions, not been able to make use of social comparison en limited opportunities for comparison. To further investigate unexpected answering behavior I studied the individual score patterns.

Person fit at the individual level

Researchers typically administer personality questionnaires because they are interested in creating profiles of trait scores that can be used to diagnose, council, or predict behavior. Additional information obtained from studying the configuration of item scores are then especially useful when similar score profiles are the result of very different configuration of item scores. Therefore, I depicted profiles of trait scores from my SPPC data in Figures 5a (boys) and 5b (girls).

A researcher would (correctly) conclude on the basis of the three profiles in Figure 4a 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 produces a very inconsistent item score pattern that consists of mostly extreme scores (SC:422124, SA:444414, AC:411444, PA:313414, BC:124443, GS:344143), whereas the patterns of children 94 and 242 consists of an alternation of 1, 2, 3, and 4 scores (person 94: SC:112422, SA:443423, AC:444322, PA:222242, BC:333333, GS:424233 ; person 242: SC:222232, SA:443333, AC:333343, PA:322233, BC:433223, GS:343333). Figure 5b depicts the score profiles of three girls with relatively low scores on the SPPC. For the patterns in Figure 5b pattern 145 is the result of inconsistent response behavior; again I observe mostly 4 and 1 scores (SC: 411211, SA:241412, AC:113112, PA:114112, BC:111234, GS:314421), whereas patterns 325 and 461 consist of mostly 2 and 3 scores (325: SC:132211, SA:221413, AC:213143, PA:241112, BC:122111, GS:212221; 461: SC:111212, SA:332122, AC:223132, PA:221222, BC:323232, GS:232322).

I inspected the 35 item score patterns in the tail of the Z_{GE} distribution with values larger than 2.0. In Table 4 I depicted the ten most aberrant response patterns in the sample, arranged according to their Z_{GE} score, where number 581 has the highest score. A general trend in these score pattern is that there is an alternation of 1 and 4 scores which is unexpected. 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), whereas he also indicates that he would like that more children liked him (item 5).

In the Appendix I report some observations during test taking. From these observations I 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. I further validated this conclusion by interviewing the teachers.

Table 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 the inconsistent answering behavior may be lack of cognitive skills to understand the questionnaire and as a result interpreting the SPPC in an idiosyncratic way.

Re-administration of the SPPC

I re-administered the SPPC to the 27 children with $Z_{GE} > 2.0$ to obtain a score that may be more representative than the score obtained at the first test administration (I could not re-administer the SPPC to all 35 children with $Z_{GE} > 2.0$ because one school did not allow me to re-administer the SPPC a second time). Again, I calculated the person-fit statistics to determine consistency of answering behavior. These Z_{GE} values from the second administration ("second") are also given in Table 5, together with the Z_{GE} values obtained at the first administration ("first"). The difference between the two Z_{GE} scores (Diff) can be found in column six in Table 5. Values around 0 and negative values indicate that the child also produced an inconsistent pattern at the second administration. The identification numbers of these person numbers are given in bold in Table 5.

As expected, the second Z_{GE} scores were lower than the first Z_{GE} scores. Two observations are important here: 8 out of the 27 children produced again irregular item score patterns. This point at some serious problem filling out the questionnaire, a tentative conclusion is that for four children (children 306, 144, 145, and 392) this is due to cognitive problems: learning retardation, 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 have

troubled home situations, difficult relations with their parents and are very insecure. This may be reflected by the way they fill out the questionnaire. It has been suggested in the literature (Harter, 1983; Leung & Leung, 1992) that environmental context besides age changes influence self-concept.

An example of a child that answered very inconsistently the first time (Z_{GE} = 4.01), but answered very consistently the second time (Z_{GE} = -.11) is child 26. The boy worked very fast and inaccurate during the first administration. The second time the SPPC was administered in smaller groups and because of group pressure, he filled in the SPPC more accurately. Another interesting example is child 393. During re-administration the author found out that this girl did not understand how to answer the questions. The 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 the author helped the girl with answering the questions again. This administration resulted in a consistent answering pattern, whereas the first administration resulted in an inconsistent pattern. Her teacher explained that this girl has problems with reading comprehension skills.

When I interpret the overall results, a general picture arises that patterns that do not fit the IRT model are especially the result of not understanding how to fill out the questionnaire, not understanding the wording due to low cognitive ability and /or lack of a consistent self-concept. (see Table 5). This is in line with findings in the literature. Cramer (1999) found a linear relation between ego development and intelligence and Marsh and Holmes (1991) reported that children who filled out the SPPC incorrectly had a lower self-concept and poorer academic performance. Thus lower level of intelligence may explain inconsistent answering behavior through (a) not

understanding the format and the wording and (b) a less developed self concept. As a result a researcher should be careful in interpreting the total scores on the subscales for these children.

Discussion

In this study I investigated the fit of individual score patterns on a self-report inventory to an IRT model and tried to interpret the results by means of additional information from observation and interviews. By means of these additional data I obtained more insight in possible reasons 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 would guarantee that identified patterns have associations with external criteria or diagnostic categories it is important to use information from other sources.

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. I found some evidence that inconsistent response behavior may be the result of a less-differentiated self concept. Changes in self-description are related to the development of cognitive abilities across childhood and adolescence (e.g., Harter, 1990). Older children are more able to evaluate the postulates of their self-concept from the standpoint of whether they are internally consistent. My results show that I can identify children with a self concept that is internally inconsistent. Although it is known from the literature that young children are inclined to give extreme responses, I showed that also old children with learning disabilities may give extreme responses. For these children, the response process is not adequately specified which may have important consequences for the diagnostic process.

Another 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. 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 person are flagged by studying the subtest score profiles. Children with extreme scores on the SPPC will fall in the 85 and 15 percentiles and will be studied in more detail by the personality researcher.

It is strange that in high-stakes psychological testing item score patterns are not screened routinely with respect to their consistency with the underlying latent trait. Although for large instruments there are response consistency scales (such as for the MMPI-2), for many scales this is not the case. I showed that using some simple indices together with information from observation and interviews can help a researcher whether he/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 using simulated data. In addition, I found that not only faking may be a problem. In the present study, some children did not understand the wording and/or had a less differentiated self concept and consequently the SPPC was not suited to measure their self-concept.

Although I tried to get some insight into the reasons why children produce atypical response patterns, it should be realized that my explanations were post-hoc. For some children I could not explain why the patterns were unexpected. This illustrates the fact that “there is and always will be a gap between the organized world of a mathematical measurement and the messy world of real people reacting to a real set of items” (Molenaar, 2004).

References

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting Item Response Theory Models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36 (4),* 523-562.

Cramer, P. (1999). Ego functions and ego development: Defense mechanisms and intelligence as predictors of ego level. *Journal of Personality, 67*(5), 735-760.

Eccles, J., Wigfield, A, Harold, R. D., & Blumenfield, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64*, 830-847.

Eder, R. A. (1990). Uncovering young children's psychological selves: Individual and developmental differences. *Child Development, 61*(3), 849-863.

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

Emons, W. H. M. (2003). *Detection and diagnosis of misfitting item-score patterns*. Unpublished doctoral dissertation.

Emons, W. H. M., Sijtsma, K., & Meijer, R.R, (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101-119.

Ferrando P.J., & Chico E. (2001).Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection *Educational and Psychological Measurement*, *61*, 997-1012.

Guttman, L. (1950). *The basis for scalogram analysis*. In S.A. Star, & J.A. Clausen (Eds.), Measurement and prediction} (pp. 60-90). Princeton: Princeton University Press

Harter, S. (1985). *Manual for the self-perception profile for children*. Denver: University of Denver.

Harter, S. (1990). Developmental differences in the nature of self-representations: Implications for the understanding, assessment, and treatment of maladaptive behavior. *Cognitive Therapy and Research, 14,* 113-142.

Harter, S., & Pike, R. (1984). The pictorial scale of perceived competence and social acceptance for young children. *Child Development, 55,* 1969-1992.

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73 (2)*, 509-527.

Junker, B. W. & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*(3), 211-220.

Marsh, H. W., Craven, R. G., & Debus, R. (1991). Self-concepts of young children 5 to 8 years of age: Measurement and multidimensional structure. *Journal of Educational Psychology, 83*(3), 377-392.

Marsh, H. W., Craven, R. G., & Debus, R. (1998). Structure, stability and development of young children's self-concepts: A multicohort-multioccasion study. *Child Development, 69*(4), 1030-1053.

Marsh, H. W., & Holmes, I. W. (1990). Multidimensional self-concepts: Construct validation of responses by children. *American Educational Research Journal, 27*, 89-117.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic.

Applied Psychological Measurement, 18, 311-314.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*, 72-87,

Meijer, R. R. & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9, 354 - 368.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.

Molenaar, I. W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica*, *58*, 1-20.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for windows. User's manual. Groningen*: The Netherlands: ProGamma.

Reise, S. P. , & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality* Assessment, 81, 93-103.

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

Rudner, L. M., Bracey, G., & Skaggs, G. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education, 9 (1),* 91-109.

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

Stipek, D. J. (1981). Children's perceptions of their own and their classmates'ability. *Journal of Educational Psychology, 73,* 404-410. (Dit artikel is niet online beschikbaar, maar is wel aanwezig in de bibliotheek.)

Stipek, D., & MacIver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development, 60,* 521-538.

Tellegen (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 621-663.

Van den Bergh, B. (1999). Jongens versus meisjes: zelf en leerkrachtbeoordeling op de CBSK en de CBSL [Boys versus girls: Self and teacher ratings compared with SPPC and TRS]. *Kind en adolescent*, *20*, 93-103.

Van den Bergh, B. R. H. & de Rycke, L (2003). Measuring the multidimensional self-concept and global self-worth of 6- to 8-year-olds. *The Journal of Genetic Psychology*, *210-225*.

Van den Bergh, B. R. H., & van Ranst, N. (1998). Self-concept in children: Equivalence of measurement and structure across gender and grade of Harter's Self-Perception Profile for Children. *Journal of Personality Assessment, 70 (3)*, 564-582.

Veerman, J.W., Straathof, M. A. E.., Treffers, Ph. D.A., van den Bergh, B.R. H., ten brink, L.T. (2004). *Competentiebelevingsschaal voor kinderen*. Lisse Harcourt Assessment.

Wichstrøm, L. (1995). Harter's Self-Perception Profile for Adolescents: Reliability, validity and evaluation of the question format. *Journal of Personality Assessment, 65*(1), 100-116.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking a personality instrument using appropriateness measurement. Applied Psychological Measurement, 20, 71-88.

Note

[1] An alternative would have been to use a multidimensional IRT model and a multidimensional person-fit statistic. I 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 applied researchers and school teachers that were involved in this project. Therefore, I used a relatively simple IRT model to get a first impression about data quality and quality of individual item score patterns. I do not expect that results would have been dramatically different when I had used more complex models, but future research may use these models to answer that question.

Table 1a

Mean, standard deviation and internal consistency reliability for each subscale of the SPPC for boys, girls and the whole sample

	M			SD			α		
Scale	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
SC	17.6	16.9	17.2	3.85	3.77	3.82	.773	.765	.769
SA	19.0	18.5	18.7	3.71	3.88	3.81	.730	.763	.750
AC	18.6	17.9	18.2	3.39	3.67	3.56	.588	.684	.647
PA	20.4	19.2	19.7	3.75	4.51	4.23	.768	.855	.828
BC	17.8	18.4	18.1	3.52	3.29	3.40	.719	.699	.710
GS	20.7	20.1	20.4	3.03	3.67	3.42	.673	.807	.762

Table 1b

Mean, standard deviation and internal consistency reliability for each subscale of the SPPC for young children, old children and the whole sample

	M			SD			α		
Scale	Young	Old	Total	Young	Old	Total	Young	Old	Total
SC	17.6	16.9	17.2	3.98	3.68	3.82	.760	.773	.769
SA	18.5	18.8	18.7	3.87	3.78	3.81	.699	.793	.750
AC	18.4	18.0	18.2	3.61	3.56	3.56	.600	.707	.647
PA	20.2	19.2	19.7	4.21	4.24	4.23	.813	.837	.828
BC	18.6	17.7	18.1	3.49	3.30	3.40	.686	.723	.710
GS	20.5	20.2	20.4	3.51	3.36	3.42	.723	.797	.762

Note. Young = children aged 7.92 through 9.92 years

Old = children aged 10 through 13 years

Table 2

H_g and H coefficients for the six subscales of the SPPC for the whole sample

	SPPC domain					
	SC	SA	AC	PA	BC	GS
Item 1	.35	.37	.24	.46	.24	.37
Item 2	.36	.45	.26	.39	.36	.36
Item 3	.41	.39	.26	.50	.27	.41
Item 4	.37	.31	.33	.53	.30	.44
Item 5	.43	.36	.22	.45	.37	.43
Item 6	.44	.34	.28	.53	.40	.32
H coefficient	.39	.37	.27	.47	.32	.39

Table 3

Proportion of boys and girls that chooses a response category

	Boys		Girls	
Response category	Young (n=119)	Old (n=140)	Young (n=147)	Old (n=194)
1	.101	.064	.097	.073
2	.141	.159	.144	.201
3	.252	.332	.259	.359
4	.507	.445	.500	.367

Table 4

Ten most aberrant response patterns

Person	Gender	Age	Z_{GE}	SPPC domain					
				SC	SA	AC	PA	BC	GS
581	boy	12.8	4.45	444111	444444	144414	144144	411111	144444
26	boy	9.7	4.01	444144	144114	444414	444444	444111	414144
12	girl	9.4	3.88	114444	344414	114114	421144	241144	114444
411	boy	8.9	3.76	311231	444141	413422	414441	433223	431114
306	girl	10.1	3.57	144444	111444	441144	411411	444414	444414
538	boy	11.2	3.48	441443	143444	112414	211432	211434	114141
137	boy	10.7	3.41	444441	444441	444441	244441	144411	444441
275	boy	8.1	3.32	422124	444414	411444	313414	124443	344143
101	boy	9.4	3.22	432131	144424	444441	121121	144123	141144
515	girl	8.0	3.07	141141	111411	414111	444144	144144	114111

Table 5

Explanation for the inconsistent answering behavior of the 35 children with Z_{GE}>2.0, also the first and second Z_{GE} values are given

No.	Gender	Age	First	Second	Diff	Explanation the first time
581	boy	12.80	4.45	n.a.		learning retardation, problems with reading comprehension skills and lexical processing speed
26	boy	9.67	4.01	-0.12	4.13	works (very) fast and inaccurate
12	girl	9.42	3.88	1.04	2.84	problems with reading comprehension skills and lexical processing speed, difference in cultural context
411	boy	8.92	3.76	1.37	2.39	concentration problems and child works fast and inaccurate
306	girl	10.08	3.57	3.29	0.28	learning retardation, problems with reading comprehension skills and lexical processing speed and difference in cultural context
538	boy	11.17	3.48	n.a.		learning retardation and difference in cultural context
137	boy	10.67	3.41	1.77	1.64	Perhaps the combination of honest and social desirable answering. Child has dyslexia
275	boy	8.08	3.32	1.88	1.44	problems with reading comprehension skills in combination with the fact that the child is ‘easy-going' and has a motivation problem
101	boy	9.42	3.22	1.54	1.68	problems with reading comprehension skills and lexical processing speed

No.	Gender	Age	First	Second	Diff	Explanation the first time
515	girl	8.00	3.07	n.a.		child has no problems with reading, but is very insecure and pessimistic caused by problems at home
38	boy	11.92	2.97	n.a.		concentration problems and child works fast and inaccurate
435	boy	9.33	2.94	.67	2.27	concentration problems and difference in cultural context
113	girl	9.83	2.82	-1.27	4.09	situation at home and child is not always honest
513	boy	8.50	2.78	n.a.		problems with reading comprehension skills
2	boy	8.08	2.66	2.55	.11	no explanation
418	girl	9.00	2.60	1.26	1.34	problems with reading comprehension skills
532	boy	9.67	2.56	n.a.		problems with reading comprehension skills and lexical processing speed
130	boy	9.92	2.53	1.57	.96	problems with reading comprehension skills and concentration problems
257	girl	8.50	2.53	1.57	.96	learning retardation, problems with reading comprehension skills and lexical processing speed
393	girl	7.92	2.47	.59	1.88	problems with reading comprehension skills and difference in cultural context
9	girl	8.25	2.44	1.29	1.15	difference in cultural context
312	girl	9.58	2.38	-.62	3.00	no explanation
360	girl	11.17	2.38	.30	2.08	learning retardation and problems with reading comprehension skills

No.	Gender	Age	First	Second	Diff	Explanation the first time
74	girl	8.08	2.31	.02	2.29	problems with reading comprehension skills and lexical processing speed and concentration Problems
144	boy	10.50	2.31	5.14	-2.83	problems with reading comprehension skills and concentration problems
40	girl	11.00	2.28	n.a.		learning retardation
25	boy	10.17	2.19	3.29	-1.10	works (very) fast and inaccurate
291	girl	8.83	2.16	-.71	2.87	learning retardation, problems with reading comprehension skills and lexical processing speed
409	girl	9.17	2.16	.39	1.77	problems with reading comprehension skills
138	boy	10.00	2.13	3.62	-1.49	home situation
45	boy	11.08	2.09	1.09	1.00	learning retardation, problems with reading comprehension skills
145	girl	9.75	2.09	2.08	.01	problems with reading comprehension skills and concentration problems
521	boy	8.83	2.09	n.a.		concentration problems
392	boy	7.92	2.09	2.47	-.38	problems with reading comprehension skills and lexical processing speed and concentration problems
91	boy	9.42	2.06	3.17	-1.11	situation at home

Note. **First = Z_{GE} score at the first test administration; Second = Z_{GE} score at the second test administration; Diff = First – Second**

Numbers in bold identify persons who produced an inconsistent pattern the second time

Figure 1. Item Response Functions for different models.

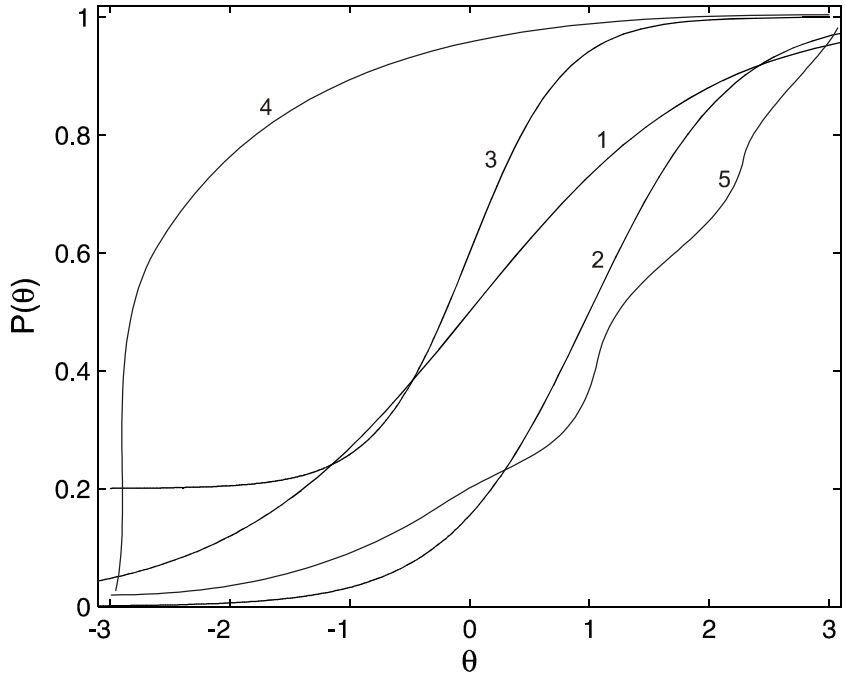
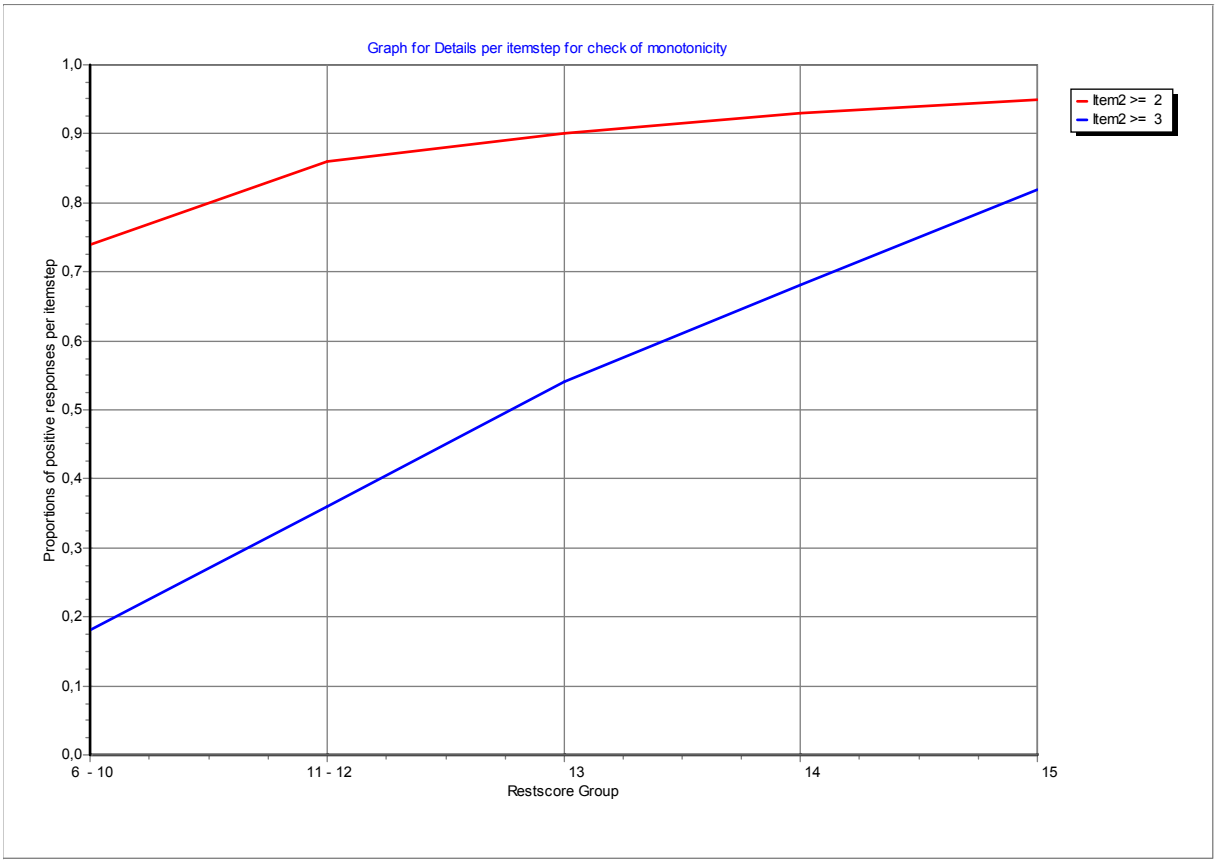


Figure 2. Example of an Item Step Response Function



Note. Restscore Group = Total score – Score at this item

Figure 3. Five Item Response Functions following the Rasch model

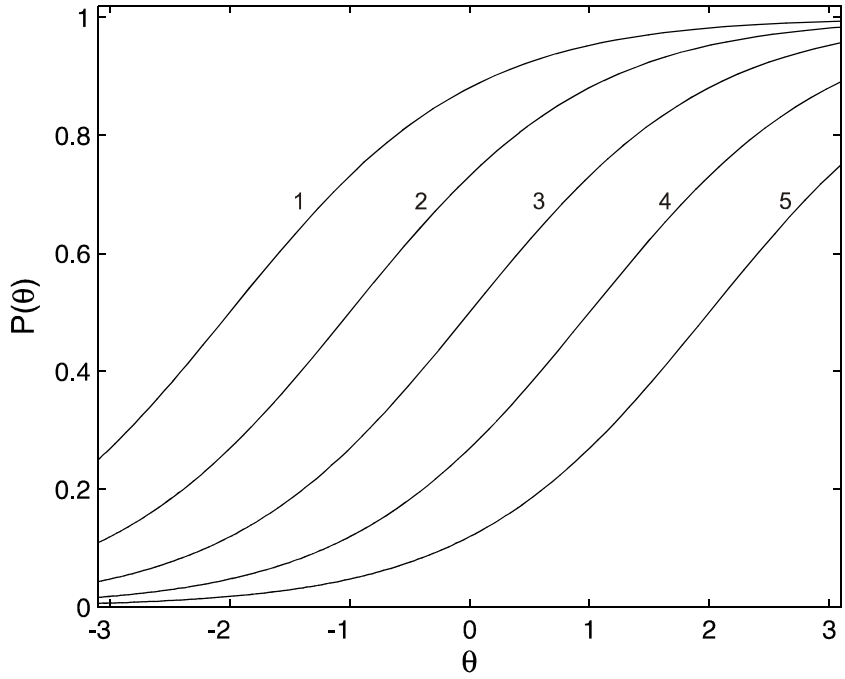


Figure 4a. Histogram of the Z_{GE} scores

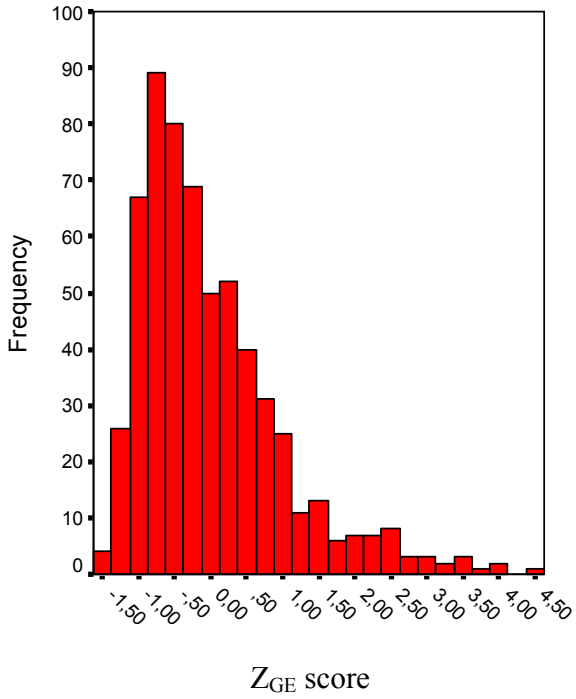


Figure 4b. Histogram of the Z_{GE} scores for young children

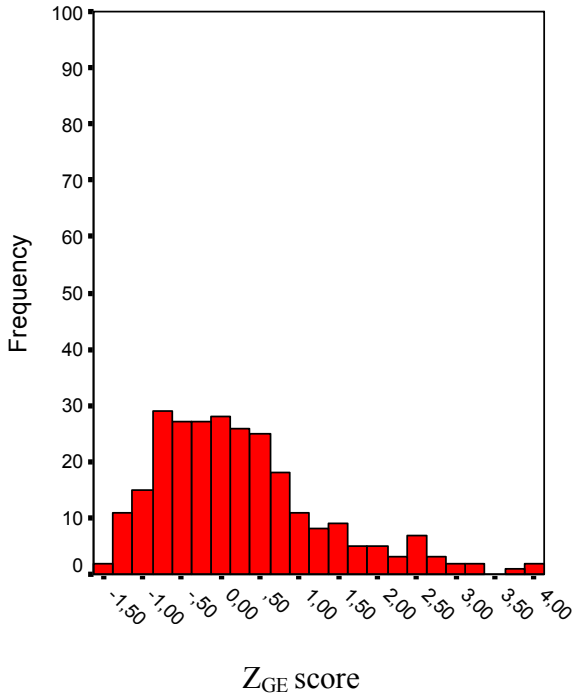


Figure 4c. Historgram of the Z_{GE} scores for old children

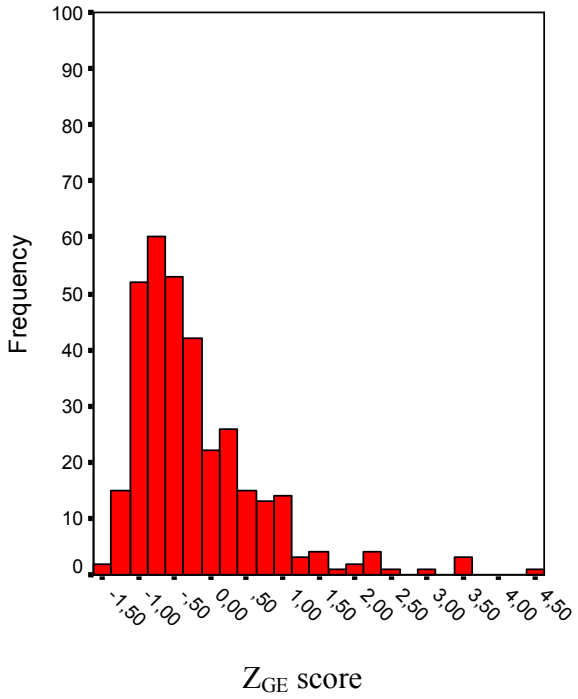
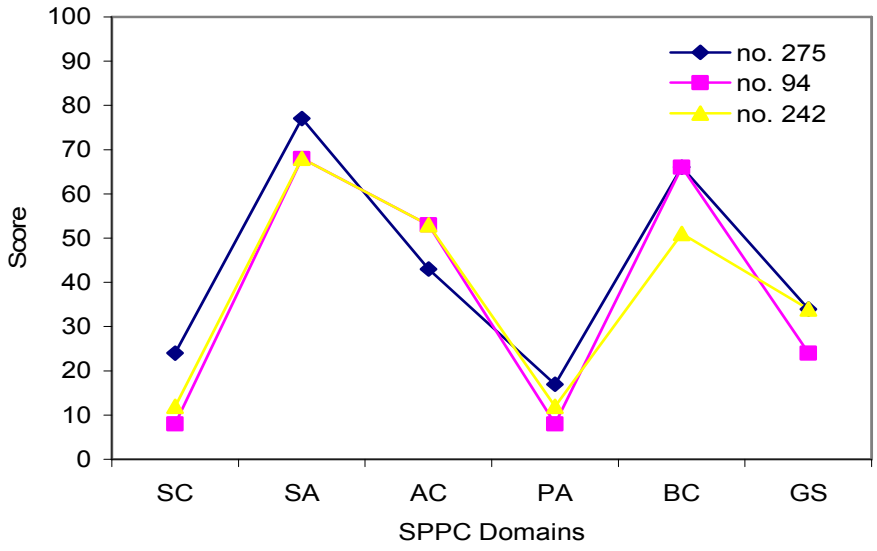
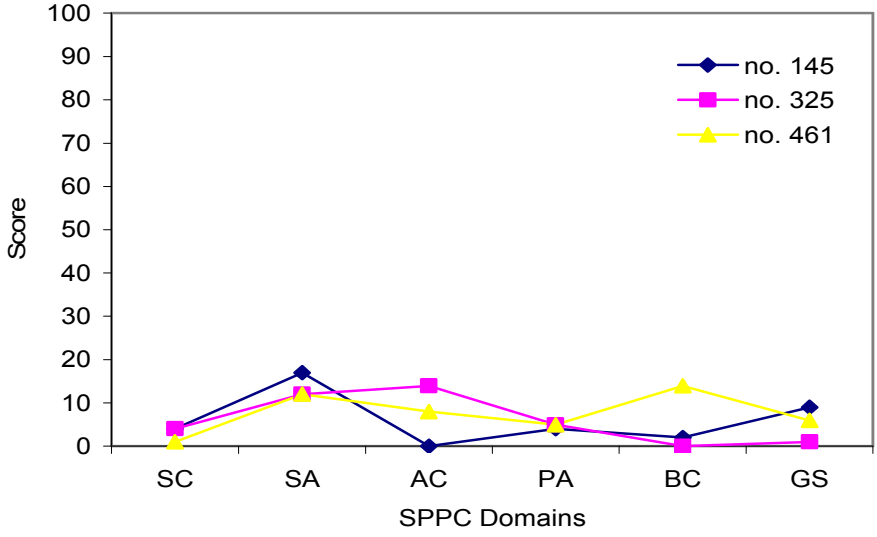


Figure 5a. Three approximately similar SPPC (score) profiles with different Z_{GE} scores



Note. Z_{GE} no. 275 = 3.32 Z_{GE} no. 94 = .27 Z_{GE} no. 242 = -1.11

Figure 5b. Three approximately similar SPPC (score) profiles with different Z_{GE} scores



Note. Z_{GE} no. 145 = 2.09 Z_{GE} no. 325 = .33 Z_{GE} no. 461 = -.45

Appendix

Observations during test administration

After instruction how to fill out the SPPC, I 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. I 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 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 like “I am good at sports”. Several children said that they did not choose this option because they did not want to make an arrogant impression. 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, I observed that young children asked more often than old children the meaning of several words. This indicates that the wording of some items was too difficult for young children and inconsistent answering behavior may be the result.