

Chapter 2

Lenient or Strict Application of IRT with an Eye on Practical Consequences

Ivo W. Molenaar

Statistics and Measurement FPPSW,
University of Groningen

Item Response Theory (IRT) is a class of measurement models that offers a variety of methods, not only to measure latent properties, but also to assess and improve the quality of such measurement. The attractive properties of IRT models, however, can only be derived when the - rather restrictive - assumptions of the model hold for the data being considered. Lenient researchers will argue that the main conclusions are fairly robust against minor or moderate violations of the assumptions. Strict researchers, on the other hand, may wish to modify either the set of items or the model until there is good agreement between data and model. This paper will argue that several strategies with a varying degree of leniency will usually lead to the same major substantive conclusions, although the results may differ in some minor aspects. This will lead to some recommendations based on the pros and cons of being lenient and being strict.

1. IRT models are great

In teaching and in consultation, many psychometricians have tried to provide motives for the use of formal measurement models, such as IRT. This use is often assumed to be beneficial for the process of scale construction and scale evaluation. It serves, for example, to identify badly performing items, that should be changed or removed from the scale. It also leads to the assessment of measurement accuracy, in the form of an estimated reliability coefficient, or still better in the form of a test information function which shows the accuracy for different levels of the latent trait being measured. Formal models also may provide person fit indices that are helpful in the detection of aberrant persons (individuals for whom it is plausible that their answers do not agree with the measurement model). In educational applications, it is often desirable to equate the scores or person parameters of students who have taken (partly) different subtests from the same domain. Sometimes one wishes to adaptively choose a next item, possibly different for different respondents, so that for each person a maximally informative test of short length is used. In other applications, it is desired to seek the best test from an existing item bank that meets a number of minimum requirements (say w.r.t. test length, content coverage and item format).

Item Response Theory has been successfully used in educational testing and in quite a few other areas. Its core virtue is probably that it assists in the statistical generalization of estimates and more general conclusions to fictitious situations like the same persons responding to other items from the same domain, other persons from the same population responding to the same items, or other test taking occasions for the same persons and the

same items. In other words, it provides at least a partial answer to the question "what would happen if we did it again". Such a reflexion about the stability of conclusions under replication is important for the bulk of empirical research in the social and behavioral sciences, although it may differ from study to study which aspects are varied under replication and which are held constant.

2. But they never fit perfectly

The possibility to reap all the nice benefits outlined in the previous section is clearly dependent on the correctness of the IRT model: performance and risks of the various strategies and procedures can only be quantified provided that the model currently considered holds for the data being examined. Unfortunately, this "ain't necessarily so", in the terminology of Sporting Life in the opera *Porgy and Bess*. Personally I go one step further and say that "it almost surely ain't so".

Most IRT models assume that there is a unidimensional latent trait, that the answers to items are locally independent for each fixed latent trait value, and that no other characteristics than someone's latent trait value influence the probability of an answer falling in a fixed response category. The more popular models also assume that the latter probabilities are logistic or normal ogive functions of the latent trait value, which only differ between items in one, two or three parameters like location and discrimination.

Arguments in favor of the plausibility of such assumptions can be based on substantive considerations (when there is some theory about the cognitive and/or emotional processes of the respondent) and on empirical considerations (although the assumptions deal with unobservable latent trait values, they have some observable consequences that provide indirect information about the fit of the model). In some cases one will have to distinguish the semantic implications of the mechanisms of randomness: Holland (1990) distinguishes the stochastic subject view, in which a fixed person may score differently on a fixed item due to short term fluctuations, from the random sampling view, in which a fixed proportion of all subjects with a fixed latent trait value scores positively on the item. For a further extension of such interpretations, see Ellis & Junker (1996). The idea that no other person characteristics than the latent trait value may influence the item response probabilities has been cast into a formal model by Ellis and Van den Wollenberg (1993).

We need not delve here into the philosophical depth of the different views on randomness, however. Regardless of our position on such issues, it is necessary to keep in mind that an IRT model and its underlying assumptions represent a Platonic ideal about measurement properties that cannot exist in real life. Therefore the characteristics of actual tests and scales are at best statistically not in conflict with a particular IRT model, but often actual test data reflect the model so badly that the validity of its assumptions is extremely doubtful.

3. Lenient or strict?

Should one be very alarmed about this sad state of affairs? The answer will vitally depend on the extent to which the main conclusions that the researcher would like to draw, are robust against violations of the IRT assumptions. If they are, one may proceed with a simple and elegant measurement model although the data tell us that this model is not, or not fully, valid. If, on the other hand, robustness is a myth, further action regarding the conflict between data

and assumptions is desirable in order to restore our trust in the measurements and in the conclusions based on them.

Such action can take different forms. If the data-model conflict is detected in a pilot study, one may try to change or replace certain items, or to change the conditions and instructions under which the test or questionnaire is administered. If the final data collection has already taken place, it is sometimes enough to remove a few misfitting items (provided that enough items are left and that the validity is not endangered by such a removal). It is far less attractive to remove a few badly fitting persons, because the measurement instrument should ideally give a latent trait value for every person in our sample. It is sometimes obvious, however, that a limited subset of persons has misunderstood the instructions or has not provided serious answers. If the model fit would dramatically improve by deleting their answer vectors and assigning them a missing value on the property that one intends to measure, that strategy may be viewed as the lesser evil. This holds in particular when, like with other occurrences of missing data, one can argue that the phenomenon is not clearly related to the main issues of the study and that the risk of biased conclusions is therefore modest.

A totally different set of strategies tackles the model, rather than the data. The fit may drastically improve by adding additional parameters for each item (like a discrimination or a guessing parameter). Sometimes, one may wish to assign different item parameters within different groups of respondents (an example is briefly discussed in section 4). More often, however, indications of such differential item functioning across subgroups of respondents are viewed as a threat to the desirable invariance of the measurement instrument across persons. In both cases, the use of extra parameters will complicate the analysis and increase the risk of chance capitalization and unstable estimation.

If evidence could be obtained that robustness holds, on the other hand, there would be no need to revise the measurement instrument or the instructions, no need to remove items or persons, and no need to add further parameters. Therefore one would like to have evidence about the stability of the main conclusions of empirical research in which IRT models are used, under variations like the following:

- with or without dubious items in the scale;
- with or without dubious persons in the scale;
- with or without additional parameters per item;
- with or without additional parameters per respondent group;
- with or without imputation of missing values.

In sections 4 and 5, some evidence of this type is presented, collected by the author in a pilot study on a limited number of data sets. It could obviously be dangerous to draw final conclusions from such a limited sample of research situations. On the other hand, the results may at least serve as illustrations of the issues addressed in this paper, and arguments will be given why other examples will often lead to similar conclusions.

4. Example 1: Organizational Climate

In a major firm offering paid services of various types, about 1000 employees from 14 job groups were asked to fill in a questionnaire about their perception of the organizational climate of their group. There were seven subscales measuring different aspects; the number of items per subscale varied between five and twelve. The items were offered with five ordered answer categories ranging from 1 (statement does not at all apply to our group) to 5 (statement does strongly apply for our group). Because the two negative categories 1 and 2 were rarely used, the items were recoded prior to analysis into new categories 0 (negative or neutral), 1 (a bit positive) and 2 (quite positive). Some subscales were not presented to all 14 job groups, which explains that the number of useful respondents varies between 399 and 905.

Three different IRT models were successfully fitted to this data set: the nonparametric Mokken (1971) model as implemented in the MSP software (Molenaar et al., 1994); the partial credit version of the Rasch model as implemented in the OPLM software (Verhelst, 1992), and the extension of this partial credit model to the One Parameter Logistic Model (Verhelst, 1992; Verhelst & Glas, 1995), which permits the discrimination parameter of the partial credit model to vary between items as long as it assumes a limited number of integer values, by which Conditional Maximum Likelihood estimation of item category parameters is still possible. The Mokken model orders respondents by their unweighted Likert score summed across items; for the other two models the bias corrected Warm (1989) estimates of the person parameters were used to locate the persons on the latent continuum measured by the subscale.

Table 1 gives the correlations per subscale between the three measures; they range from 0.85 to 1.00. Generally speaking the longer subscales have higher correlations. For inferences at the group level, such as score differences according to gender, or correlation of score with age, it appears that each of the three IRT models will lead to similar conclusions. A detailed inspection reveals, however, that in a few exceptional cases respondents are differently ordered by the three models.

<i>scale</i>	<i>k</i>	<i>n</i>	$r(T,U)$	$r(T,W)$	$r(U,W)$
A	5	905	0.89	0.85	0.96
B	5	399	0.90	0.86	0.94
C	6	855	0.91	0.91	0.99
D	6	635	0.94	0.91	0.97
E	7	747	0.91	0.91	0.99
F	12	871	0.95	0.94	1.00

Table 1: Subscale length *k*, number of valid responses *n*, and correlations between unweighted sum score *T*, Warm estimate *U* from the partial credit model, and Warm estimate *W* from the OPLM model

One may expect that such high correlations will be a regular finding. The Warm estimate *U* is a strictly increasing function, say $h(T)$, of the sum score *T*. Although it is non-linear, only a few respondents have (almost) zero or perfect scores, and most have *T*-values in an interval in which $h(T)$ is almost linearly related to *T*, the more so when the scale is longer. For an explanation of the still higher correlations between *U* and *W*, observe that the correlation between the simple sum *T* and the weighted sum $T^* = \sum a_i x_i$ **Fehler! Schalterargument nicht angegeben.** tends to be very high for all nonnegative weights a_i . In the present case the

weights used to fit the OPLM model are vectors like 1,3,2,2,2,1 or 5,4,6,3,8,6, which leads to almost perfect linearity between T and T*. Computation of the Warm estimates U and W for both models will hardly disturb this relation, although each model has its own item category parameters. By composition, the relations between T and W are only slightly lower than those between T and U.

Another aspect studied for the present data is the introduction of subgroup specific item parameters. In all three models for subscale E, it was clear that item 5 did not fit very well. Both from the CHECK=GROUPS option in MSP and from detailed output for OPLM, it was concluded that the fit would greatly improve if the two item category parameters of item 5 were allowed to be different in job group 9 from what they were in the other job groups. Although item parameter invariance across subgroups is generally viewed as desirable, it is not really surprising that there may be a job group in the organization where one specific aspect is judged very unfavorably by most respondents, relative to the other aspects of the same concept.

After this modification, the partial credit model had a very good fit. Although the probability of a positive answer in job group 9 now had a location shifted by 2.8 logits compared to its position for the other job groups, the effect on the Warm estimates U was small: depending on their position on the continuum and their score obtained on item 5, respondents' person parameter estimate would change by 0.3 to 0.4 logits as compared to the corresponding estimate in the model without a group specific position of item 5.

The subscale scores discussed here were used to (a) fit a LISREL model describing the relations between the subscales, and (b) study the relations between quality aspects as perceived by the personnel to similar aspects as perceived by the clients of the organization. For neither of those, the variations described above (three IRT models, introduction of group specific item parameters for one item out of seven in order to improve model fit) made a major difference.

5. Example 2: Bedside Neuropsychology

For a group of patients hospitalized after a cerebrovascular accident, a neurologist tried to obtain scores on ten 4-point items measuring problems in simple neuropsychological tasks in the sensory and motor domain. Due to the condition of the patients and various medical and organizational problems, only 49 out of the 128 patients have valid scores on all ten tasks; 66 other patients have one to four missing scores. The remaining 13 patients have more than four missing task scores.

Among the main aspects that were of interest to the neurologist, we mention the following:

- a) do the ten tasks measure a unidimensional latent construct, and if not, are there subsets satisfying unidimensionality?
- b) to what extent can the (sub)scale values of the patients be explained by four predictors, described below, and their interactions?

In the process of finding answers to these questions, one may once more assume a lenient or a strict attitude about the validity of answers obtained by IRT models, as will be illustrated below. The frequent occurrence of missing data in the present example requires special attention, however. As a first step, it will be investigated whether the pattern of missingness

can be viewed as random. Afterwards, the discussion returns to the impact of model decisions on the answers to questions a) and b).

In several respects, the distribution of the 222 missing values across the 128*10 data matrix is not random. First, there is no homogeneity across items: tasks 2, 4 and 5 have more missing observations than the others (see Table 2, two-sided binomial tests of success probability 222/1280; the tests are not independent, see below):

<i>nr</i>	<i>name</i>	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>9</i>	<i>fraction missing</i>	<i>two-sided P</i>
1	ZOEK	88	0	3	17	20	p1=20/128=0.156	0.70
2	VAST	71	4	2	12	39	p2=39/128=0.305	0.0004
3	LSTOT	82	13	16	1	16	p3=16/128=0.125	0.18
4	MOTEX	66	0	7	7	48	p4=48/128=0.375	< 0.0001
5	PARHYP	76	0	10	3	39	p5=39/128=0.305	0.0004
6	ANGN	76	0	9	29	14	p6=14/128=0.109	0.062
7	ACUST	74	6	24	10	14	p7=14/128=0.109	0.062
8	TACT	62	17	19	23	7	p8= 7/128=0.055	0.0001
9	VIS	60	8	15	31	14	p9=14/128=0.109	0.062
10	IGCH	44	0	9	64	11	p10=11/128=0.086	0.0074

Table 2: Observed marginal frequency distributions

In Table 3 the 1024 patterns are grouped according to number of missing tasks per person.

# missing	0	1	2	3	4	5	6	7	8	9	10	total
frequency	49	26	20	15	5	3	1	5	1	3	0	128

Table 3: Frequency distribution of number of missing tasks

From the null model in which $P(\text{task } i \text{ is missing}) = p_i$ is taken from Table 2, but missingness is independent across tasks, it is found that one expects 17.4 cases with 0 missing and 40.5 cases with 1 missing. At the other end, one expects only .0008 cases with 8 missing, .00004 cases with 9 missing, and .0000005 cases with 10 missing. So there are too many persons with 0 missing and too many with 7,8,9 missing, contradicting independence of missingness across tasks. This confirms the information that missing task scores were often related to poor condition of the patient. Further analysis of pairwise missing tasks, not reported here, shows that for all 45 task pairs, the number of persons with both tasks missing exceeds the frequency predicted under independence given the marginal frequency p_i of missingness. For 35 pairs the result is significant at a two-sided 0.05 level.

<i># missing</i>	<i>frequency</i>	<i>average valid score</i>
0	49	0.50
1	24	0.63
2	20	1.29
3-4	19	1.87
5-9	13	2.01

Table 4: Average valid score as a function of number of missing tasks

The last and perhaps most important aspect of missingness to be discussed is: does the fact that a score is missing on one task alter the frequency distribution across the valid scores 0,1,2,3 on another task? Table 4 provides a clear answer: missing observations on some tasks and average valid score on the remaining tasks are positively associated.

The existence of 17 percent non-randomly missing values implies that some caution is required in the investigation of the effect of leniency or strictness on the two major goals a) and b), see above, for which the data were used by the neurologist. For question a), the existence of one or more unidimensional scales, it was decided to use the nonparametric IRT model of Mokken (1971) both in its polytomous and in its dichotomous form, with listwise deletion of missing values. This is the standard option in the MSP software package, see Molenaar et al. (1994, chapter 5) for a discussion. Because only 49 out of the 128 patients have valid scores on all ten tasks, it was decided to add an analysis in which tasks 2, 4 and 5 were omitted; as shown in Table 2 these three tasks have far more missing values than the others. In the analysis for seven tasks 82 subjects are left after listwise deletion.

#patients	#tasks	categ.	<i>H</i>	<i>Rho</i>	score range	mean	st.dev.	skewn.
49	10	0-3	0.82	0.94	21*0...1*25	4.7	6.8	1.7
49	10	0-1	0.86	0.94	21*0...1*10	2.1	2.8	1.4
82	7	0-3	0.85	0.94	34*0...1*21	5.4	6.6	1.0
82	7	0-1	0.87	0.93	34*0...10*7	2.2	2.6	0.7
28	10	0-3	0.75	0.92	1*1...1*25	8.3	7.2	1.2

Table 5: Mokken analysis of the bedside tasks

	T ₇	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
T ₁	0	21																						
	1		1																					
	2			4																				
	3				5																			
	4					3																		
	5						2																	
	6																							
	7								3															
	8									1														
	9										2													
	10																							
	11												1											
	12													2										
	13																							
	14																							
	15																							
	16																	1						
	17																							
	18																							
	19																							
	20																							
	21																							
	22																							
	23																							
	24																							
	25																		1			1		
	26																							
	27																							
	28																							
	29																							
	30																							

Table 6: Cross tabulation of the sumscores T₁₀ (all 10 items) and T₇ (subset of 7 items) each scored 0 to 3, Neuropsychology data, 49 complete cases.

Table 5 shows that there is a strong unidimensional Mokken scale in all variants of the analysis. This follows from Loevinger's H coefficient of scalability and from the estimated reliability Rho, both of which assume very high values. The first four lines refer to the combination of 10 or 7 items and 4 or 2 scale categories. The last one analyzes ten items after omission of all patients with zero scores (good performance) on all ten tasks. It was added in order to check whether the favorable results could be an artifact of having 21 out of the 49 useful subjects with only zero scores. Although H and Rho are slightly lower for the remaining 28 patients, the scale remains very good.

In all cases, there was no reason at all to suspect that some items would measure a different dimension or not fit into the model (for reasons of space, these results are not presented).

This raises the question whether the use of the sumscores of seven rather than ten tasks would have made much difference for ordering the patients from low (no adverse effects) to high (many problems with the tasks). The answer is presented in Table 6 in the form of a cross tabulation of the two sum scores T7 and T10 for the 48 complete cases. For all 40 cases with $T10 < 10$ one has $T7 = T10$ and the ordering remains the same. For the cases with a higher total score, the results are closer to a linear relation $T7 = 0.7 * T10$.

It should not be viewed as surprising that a sumscores on a test of length ten correlates highly with a sumscores obtained after omitting three items. Suppose, for simplicity, that all items have the same standard deviation, and that the correlation matrix exhibits only three different values: ρ_7 , between each pair from the remaining seven items, ρ_3 , between each pair of the three omitted items, and ρ_{ob} , for each of the 21 pairs consisting of one omitted and one retained item.

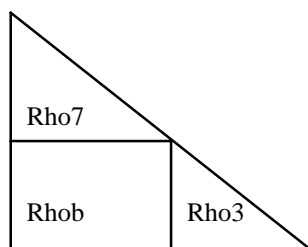


Figure 1: Explanation of ρ_7 , ρ_3 and ρ_{ob}

ρ_7	ρ_{ob}	ρ_3	$\rho(T10, T7)$
.0	.0	.0	.84
.2	.0	.0	.91
.4	.0	.0	.94
.2	.0	.2	.89
.4	.0	.2	.92
.4	.0	.4	.90
.2	.2	.2	.94
.4	.2	.2	.95
.4	.4	.4	.97

Table 7: Correlation between the sumscores T10 and T7 for various values of the inter-item correlations ρ_7 , ρ_{ob} and ρ_3 , see text

Table 7 shows the effects on $\rho(T10, T7)$ for some possible combinations of three different inter-item correlations, 0.0, 0.2, and 0.4. Table 7 shows that T7 and T10 would

correlate over .90 for the realistic values $\rho_7=.2$ and $\rho_7=.4$ even if the three removed items were purely random noise, i.e. correlated zero between themselves and with the seven remaining items.

In the further analysis of the neurological bedside data, the very low number of complete cases made it desirable to impute a total score for incomplete cases before passing to the regression analysis. The strong scalability result inspires some trust in an effort to use the valid data to impute the missing ones; the 13 cases with more than four missing task scores, however, will be omitted because their imputed total score might well be very unreliable.

	T ₇	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
T ₁	0	37																					
	1		4																				
	2		2	5																			
	3				8																		
	4					7																	
	5						3																
	6																						
	7					1	1		3														
	8								1	2													
	9									1	2												
	10																						
	11								1				3										
	12													3		1							
	13																						
	14																						
	15											1	2										
	16													1	1								
	17													1	1								
	18																1						
	19														1								
	20																	1					
	21																		1				
	22																		1	1			
	23																			1			
	24																	2			1		
	25																		1			1	
	26																						
	27																				3		1
	28																				2		1
	29																						4
	30																						

Table 8: Cross tabulation of the sumscores T₁₀ (all 10 items) and T₇ (subset of 7 items) each scored 0 to 3, Neuropsychology data, 115 cases, after EXP imputation.

For the other 115 cases, four imputation methods were considered:

- EXP: Expert judgement by the neurologist;
- MVAL: inserting the mean of the patient on the valid task scores;
- MEAN: inserting the mean of the estimated cell distribution;
- MODE: inserting the mode of the estimated cell distribution.

For methods MEAN and MODE the OPMISS software (Nap, 1994) was used, in which the probability of a score 0,1,2,3 is estimated for each missing cell from the item category parameters and the person parameters as estimated by OPLM (Verhelst, 1992). The often

applied strategy of inserting the mean task score of the other patients was not considered, because the results in Table 4 indicate that this might lead to serious underestimation.

In Table 8 the sumscores T7 and T10 are again compared, but now for 115 cases after the EXP imputation. The conclusions from Table 6 about two almost linear relations and a few order reversals are confirmed. This also holds for the other imputation methods, for which the results will not be presented.

Table 9 gives a comparison of the regression results. The sum score of a patient on all ten tasks scored 0-3 (available for 49 cases and imputed for 65 others with 1 to 4 missing values) is predicted by four predictors:

LRHEM, left or right hemisphere;
 LSIZE, logarithm of the diameter of the damaged area;
 ATRO, presence or absence of atrophy;
 INFBL, infarct or bleeding,
 plus the interactions of the last three with LRHEM.

<i>Result</i>	<i>EXP</i>	<i>MVAL</i>	<i>MEAN</i>	<i>MODE</i>
Multiple R	.827	.824	.823	.831
Adj.Rsquared	.662	.657	.657	.670
constant	-3.1 (2.9)	-3.6 (3.0)	-3.5 (3.0)	-3.0 (2.9)
LRHEM	-3.4 (3.9)	-3.2 (4.0)	-3.1 (3.9)	-3.7 (3.9)
LSIZE	1.2 (0.6)	1.5 (0.6)	1.4 (0.6)	1.1 (0.6)
ATRO	0.8 (0.5)	0.9 (0.5)	0.9 (0.5)	0.8 (0.5)
INFBL	1.7 (2.1)	1.9 (2.2)	1.9 (2.1)	1.7 (2.1)
LRHEM*LSIZE	2.9 (0.8)	2.7 (0.8)	2.9 (0.8)	3.2 (0.8)
LRHEM*ATRO	0.6 (0.6)	0.4 (0.6)	0.5 (0.6)	0.5 (0.6)
LRHEM*INFBL	4.1 (2.8)	4.7 (2.9)	4.2 (2.9)	4.5 (2.9)

Table 9: Regression results for four imputation methods

The table shows that some regression coefficients are somewhat different when the dependent variable has been differently imputed. Even for the first and the last coefficient, however, the differences are small, both in terms of the effect on the prediction and in terms of their standard error. The overall impression is that the influence of the strategy choice is very modest. In a stepwise analysis not reported here, only three predictors remained which explained 65 percent of the variance, and that result again did not change with the imputation method.

6. So what? Some conclusions

It has been argued that the main conclusions from the two datasets described above seem to be rather robust against several variations in the choice of an IRT model or in the choice of a strategy to handle certain deficiencies in the data. At the level of single individuals, however, such choices can make quite a difference.

The skeptical reader may well react by "So what"? A few arguments have been given for the plausibility of obtaining similar findings in many other data sets: Both after a somewhat non-linear transformation (like from total score to estimated ability on a logit scale) and after linear combination with different weights (including omission of a few items from a scale),

the correlation with the original person scores will be very high. There will always be exceptions to this footrule, however, and moreover it is not clear how a substantive researcher or a consulted psychometrician should react to this message that it will probably not matter too much. It appears appropriate to end this paper with a few tentative conclusions and recommendations.

Although person parameters estimated from a logistic test model have many advantages in the situations sketched in section 1, the present author proposes to give three cheers to the raw total score. It performed quite well in our examples, it is easy to use for a substantive researcher and easy to understand for the readers of his/her research reports, and unless there are clear counter-indications it may be expected to produce the same major substantive conclusions as more sophisticated measures of a latent property. In the Mokken model it is used to order the persons, and in the Rasch model it is the sufficient statistic for inferences about the person parameter on the logit scale. So if we take total score seriously we are in good company when such a model fits in data; whether it is worth to use total score as an ordinal measure only, or to replace it by a person parameter derived from a parametric IRT model, will depend strongly on the data and the research goal.

A second aspect is the difference between a myopic and a helicopter view of the data and the analysis. In the wealth of detail provided by modern computers, it is easy to find a few cases for which two models or two strategies lead to different conclusions, a few regression coefficients that differ non-negligibly, or a few items that fit under one model but not under the other. In both examples that were discussed, however, the majority of cases and certainly the main conclusions drawn from the study remained almost unaffected under our alternative research strategies. Moreover, reasons have been given why it is plausible that these findings will hold for other examples too.

A third conclusion is that robustness of the kind described above has been too little explored. It is known from other fields that one cannot simply say that a statistical procedure "is robust". Whether one deals with a simple two-sample t-test or with maximum likelihood estimation of a vector of fifty parameters in a LISREL model, the best one can hope to find in a general robustness study is a tentative conclusion that the procedure appears to be affected by not more than $x\%$ (in level, power, bias, standard error) if assumption A is violated by no more than $z\%$, if the other assumptions hold, and and if sample size is at least N .

Given modern computing power, it could be a good idea to put less effort into robustness studies aiming at general conclusions, but to recommend that each researcher reports a few variations of his/her own data analysis. Example: the conclusion that students in program A score about 10 percent higher than those in program B was based on the OPLM model with unequal discriminations, but a further analysis shows that it remains valid under the dichotomous Rasch model and under the non-parametric Mokken model. Or: for reasons of validity it was decided to keep the two items with a dubious fit to the Rasch model in the scale, but a re-analysis without them led to regression coefficients which differed no more than 15% from those reported below. Or: the asymptotic standard errors in table X cannot be fully trusted with our sample size of 115, but the values obtained from 50 bootstrap samples were not very different.

In a few cases, the poor researcher would have to report that some seemingly innocent variations made a great deal of difference. Then, however, it is only fair to warn the reader

about this. Moreover, it then becomes very important to give a convincing defense of the model and method on which the major conclusions are based.

Hopefully, there will be many cases where a simple analysis (easily understood by most readers) and a more sophisticated one (which makes the more critical readers happy) lead to similar conclusions. Especially in the case of small sample sizes the sophisticated procedures may be more vulnerable than the simple ones; when there are enough data there are more reasons to prefer a more elaborate model because it can incorporate more important features and its chances to survive cross validation grow rapidly with sample size.

One thing must be clear: the high correlations and robust conclusions displayed earlier in this paper are by no means a good reason for sloppy modeling or sloppy measurement. Careful modeling and careful measurement bear some resemblance to airbags and safety belts in a car: they are somewhat costly, and one rarely needs them, but when one does, their presence matters an awful lot!

Acknowledgement

The author is grateful to John de Jong (CITO) and to two reviewers for their comments on an earlier version of this paper.

References

- Ellis, J.L. & Van den Wollenberg, A.L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika*, 58, 417-429.
- Ellis, J.L. & Junker, B.W. (1996). Tail-measurability in monotone latent variable models. *Psychometrika*, to appear.
- Mokken, R.J. (1971). *A theory and procedure of scale analysis*. New York/Berlin: Walter de Gruyter/Mouton.
- Molenaar, I.W., Debets, P., Sijtsma, K. & Hemker, B.T. (1994). *User's Manual MSP*. Groningen: iec ProGAMMA.
- Nap, R.E. (1994). OPMISS: Handling missing data in OPLM. *Heymans Bulletin* (internal publication) HB-94-1173-IN. Groningen: Department of Statistics and Measurement Theory, University of Groningen.
- Verhelst, N.D. (1992). *Het eenparameter logistische model (OPLM), OPLM-manual*. OPD Memorandum 92-3. Arnhem: Cito.
- Verhelst, N.D. & Glas, C.A.W. (1995). The one parameter logistic model. In: G.H. Fischer & I.W. Molenaar (eds.), *Rasch models: foundations, recent developments and applications*, 215-237. New York: Springer.
- Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.