Quality & Quantity **34:** 177–191, 2000.

© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Latent Class Analysis of Respondent Scalability

G. VAN DEN WITTENBOER[1,⋆], J. J. HOX[2] and E. D. DE LEEUW[3]

[1]*Department of Education, University of Amsterdam;* [2]*Utrecht University;* [3]*MethodikA, Amsterdam*

Abstract. The psychometric literature contains many indices to detect aberrant respondents. A different, promising approach is using ordered latent class analysis with the goal to distinguish latent classes of respondents that are scalable, from latent classes of respondents that are not scalable (i.e., aberrant) according to the scaling model adopted. This article examines seven Latent Class models for a cumulative scale. A simulation study was performed to study the efficacy of different models for data that follow the scale model perfectly. A second simulation study was performed to study how well these models detect aberrant respondents.

Key words: latent class analysis, person fit research, measurement error, respondent error.

1. Introduction

Four well-known sources of measurement error in surveys are the questionnaire (e.g., question wording), the data collection mode (e.g., face to face or telephone interviews), the interviewer, and the respondent (Groves, 1989). Unlike the first three sources of measurement error (questionnaire, mode, interviewer), the fourth error source (respondent) is difficult to minimize. Respondents can be instructed in what is expected from them (e.g., think carefully, use the answer categories provided) and they may be motivated to do their best. But it is difficult to manipulate a respondent to reduce the respondent error. Therefore, research on respondent errors has concentrated on attempts to identify those respondents who produce errors and search for their unique properties.

An important problem in this type of research is how to measure respondent error. Groves (1989: 445–446) summarizes this as follows: “Measurement errors are generally viewed as specific to a particular measure, one question posed to the respondent. Only by identifying response tendencies of respondents over many questions can inference about respondent influences on measurement error be made. Then only by comparing different respondents on the same task can characteristics of the respondents which produce measurement error be identified”.

One promising approach is the application of person fit indices to detect inconsistent respondents (Meijer and De Leeuw, 1993; Meijer, 1994; De Leeuw and Hox,

⋆ Author for correspondence: Department of Education, University of Amsterdam, Wibautstraat 4, NL 1091 GM Amsterdam, The Netherlands, phone: + 31 20 5251529; fax: +31 20 5251200; e-mail: witten@educ.uva.nl

1994). A very different approach is the application of ordered latent class analysis with the goal to distinguish latent classes of respondents that are consistent in their responses and latent classes of respondents that are not consistent (i.e., aberrant). In both approaches, a psychometric scaling model (e.g., the Guttman model) serves as a benchmark for aberrance. (See Forman, 1988, for a discussion of latent class models for nonmonotone items).

In this paper, we investigate whether latent class analysis is a useful tool in the study of respondent error. We start with a short evaluation of current approaches in person fit research. This is followed by an overview of a number of latent class models for Guttman type data (Section 3). To investigate how well these models perform, we set up a simulation study which is described in Section 4. The simulation study has two goals: (1) identifying models that provide a satisfactory fit to a highly scalable but not perfect item set, and (2) that are capable of identifying aberrant respondents in a second 'polluted' data set. We end with a discussion of the effectiveness of ordered latent class analysis for the detection of aberrant respondents (Section 5).

2. Person Fit Research

Person fit analysis investigates whether a person exhibits response behavior that deviates from the behavior predicted by a measurement model, or from the response behavior of the majority in the population to which that person belongs. For example, if a student answers eight out of 10 questions correctly, one expects that s/he will have missed the two most difficult questions. If the two easiest questions are answered incorrectly, this response pattern is completely unexpected. To investigate response patterns, data are needed on a test or scale that consists of a number of questions about the same topic (e.g., a test, an attitude or personality scale). Furthermore, the test or scale should have good psychometric properties, that is, a high reliability and good scalability.

For persons detected as aberrant, the total scale score does not adequately reflect the attribute that is measured. For instance, using person fit indices, Levine and Rubin (1979) discuss person fit indices for the detection of cheating on aptitude tests. Harnisch and Linn (1981) use person fit indices to differentiate schools with special curriculum on math and reading. Tatsuoka and Tatsuoka (1982, 1983) identify students who use a wrong algorithm in problem solving tasks.

Although person fit indices have been developed in psychological and educational testing, applications can be found in sociology and survey research as well. Van der Flier (1980) uses person fit indices in intercultural research, Van Tilburg and De Leeuw (1991) apply person fit indices in a comparison of different data collection methods, and Meijer and De Leeuw (1993) use person fit indices to investigate aberrant respondents in a general survey.

Three groups of person fit indices can be distinguished. The first group is based on the assumptions of parametric IRT-models, such as the Rasch model (c.f. Tarnai and Rost, 1990; Molenaar and Hoytink, 1990). The second group is based on a nonparametric Item Response Theory model, such as the Mokken model (cf. Van de Flier, 1980; Sijtsma, 1988). The third group evaluates a response pattern using statistics based on the group to which a person belongs (e.g., proportion correct of items, cf. Harnisch and Linn, 1981; Meijer, 1994.) Often, some kind of deviation from the perfect Guttman pattern is used as a criterion for aberrance. Scales that fit the strict assumptions of IRT-models such as the Rasch-model are scarce, especially in survey research. Consequently, nonparametric person fit indices are more commonly used in empirical applications. One of these nonparametric person fit indices, the index ***Q*** developed by Van der Flier (1980, 1982), will be used as a benchmark in our simulation study in section 4.

Person fit indices have the advantage that they are easy to compute and can be added simply to the original data files. They have two important drawbacks. First, criteria for diagnosing aberrance are not clear. Sometimes rules of thumb exist (cf. Harnisch and Linn, 1981); sometimes a statistical test is used (Van der Flier, 1980). Using a statistical test provides a formal criterion, but since many tests must be performed (one for each respondent), this gives rise to the well-known type I error. Even if there are no aberrant respondents, still a certain number will be detected. Second, person fit indices divide respondents into two classes: 'normal' and 'aberrant', and they do not distinguish between different types of aberrant respondents.

One interesting theoretical distinction that can be made, is the distinction between aberrant response patterns that are the result of a random response process, and aberrant response patterns that are the result of a systematic but erroneous response process. For instance, in educational research one often distinguishes between 'guessers and 'cheaters'. When a student does not know the answer to some difficult questions, s/he can either guess (a random response process), or cheat by looking up the answers or looking at a neighbor's answers (a systematic response process). Both strategies may result in an aberrant response pattern, but regular person fit indices do not distinguish between the two. With measures such as attitude scales, 'guessing' could refer to respondents who answer without much thinking about the precise content of the question, a phenomenon known in survey methodology as the 'top-of-the-head' response. 'Cheating' could refer to respondents that show unexpected response behavior on extreme questions, which can be the result of a social desirability bias that shows up only on extreme questions. Because 'guessing' and 'cheating' most aptly refers to educational test data, we will use the more general terms 'random aberrant' and 'systematic aberrant' response patterns.

In the next section, we will present seven latent class models for cumulative scales of the Guttman type, which allow additional latent classes for unscalable respondents. In a simulation study, we will investigate how well these models fit a well-behaved data set, and how well they distinguish random aberrant response patterns (guessing) and systematic aberrant response patterns (cheating).

3. Latent Class Models for Cumulative Scales

Latent class analysis has a long tradition in analyzing dichotomous responses with a Guttman-type structure for the underlying scale (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968; Goodman, 1974; McCutcheon, 1987; Langeheine, 1988; Clogg, 1988). Different latent class models have been proposed for other IRT models, such as the Rasch model (cf. Rost and Langeheine, 1997). In both approaches, there may be unscalable respondents (Goodman, 1975; Clogg and Sawyer, 1981; Rost, 1990), who in fact are respondents with aberrant response patterns according to some restricted latent class model.

In this article, we confine ourselves to dichotomous probabilistic Guttman-type data as exemplified by the Mokken model (Mokken, 1971), because validated Rasch scales are rare in survey research. The Guttman cumulative scale model assumes that we have dichotomous items that are scored 'correct' (c.q. positive, yes) versus 'incorrect' (c.q. negative, no). The items can be ordered on an underlying dimension as to difficulty, and the respondents can be ordered on the same underlying dimension as to ability. The original Guttman model is deterministic; if a respondent encounters an item with a difficulty below the respondent's ability, the response will be correct with probability one. If a respondent encounters an item with a difficulty above the respondent's ability, the response will be incorrect with probability one. The Mokken model adds a probabilistic component to the Guttman cumulative scale (cf. Sijtsma, 1988).

The models we consider first are an elaboration of the deterministic Guttman model in terms of latent class analysis. These models are probabilistic in the sense that they estimate probabilities for specific responses, instead of assigning one of the two possible responses with probability one. However, they are restrictive as well, because they assume that respondents perfectly follow a cumulative scale. Idiosyncratic response behavior can only be incorporated by assuming additional latent classes for respondents that are unscalable. All these models are related to Lazarsfeld's latent distance model, and they are well known in the literature (Clogg and Sawyer, 1981; McCutcheon, 1987; Langeheine, 1988; Heinen, 1993). Next, we propose two new models that directly incorporate the possibility of unscalable response behavior. These are mixed models, because they contain both a completely deterministic part that represents the Guttman model, and a freely estimated part that represents idiosyncratic response behavior. All models are briefly discussed below.

The Latent Distance Model. The LAtent Distance (LAD) model developed by Lazarsfeld (1950) is a general model that translates the deterministic Gutmann model into a latent class model. The other models discussed here are derived from the LAD by placing additional restrictions on the parameters of the LAD model. With k items, the LAD has $k + 1$ latent classes, one for each possible score. In the following, we will formulate models for a scale that consists of seven items; the simulations presented later will also use a seven-item scale. The main ideas of

Table I. The latent distance model (the LAD model)

Latent classes	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Class 1	a_1	b_1	d_1	f_1	h_1	j_1	m_1
Class 2	a_1	b_1	d_1	f_1	h_1	j_1	$1\text{-}m_1$
Class 3	a_1	b_1	d_1	f_1	h_1	k_1	$1\text{-}m_1$
Class 4	a_1	b_1	d_1	f_1	i_1	k_1	$1\text{-}m_1$
Class 5	a_1	b_1	d_1	g_1	i_1	k_1	$1\text{-}m_1$
Class 6	a_1	b_1	e_1	g_1	i_1	k_1	$1\text{-}m_1$
Class 7	a_1	c_1	e_1	g_1	i_1	k_1	$1\text{-}m_1$
Class 8	$1\text{-}a_1$	c_1	e_1	g_1	i_1	k_1	$1\text{-}m_1$

the LAD model are presented in Table 1. In Table 1, and in the subsequent tables describing our models, the items are placed in their order of difficulty. Response category 1 refers to the correct response, and response category 2 to the incorrect response. Table 1 shows the *conditional probabilities,* given the latent class, of observing a correct, or positive answer to each of seven dichotomous items that are ordered from ‘easy’ to ‘difficult’. Cell (3,2), for example, contains the probability $b_1$that someone who belongs to latent class 3 answers item 2 correctly. The probability of observing an incorrect answer is 1-b_1, which is not in the table. The difference between these conditional probabilities and the corresponding conditional probabilities of 1 and 0 in the deterministic Guttman scale is called the ‘error rate’. Thus, the error rate of cell (1,4) is 1-f_1, because under the deterministic Guttman model the probability in that cell is equal to one, and this error rate indicates the false negatives. Likewise, the error rate of cell (8.4) is g_1, because under the perfect Guttman model this probability is equal to zero, and this error rate indicates the false positives.

Characteristic of the LAD model is that the conditional probabilities for the items 2 to 6 have the deterministic pattern. For each item, the probabilities are restricted to be equal across the latent classes for the zeroes, respectively ones, in the corresponding deterministic pattern. Exceptions are found at the end points of the scale. To avoid identification problems, the error rates of items at the end points must be set equal in each class (Lazarsfeld and Henry, 1968).

The Equal Item-Specific Error Rates Model. The Equal Item-Specific Error Rates (EISER) model assumes equal error rates for each item across the classes (Lazarsfeld and Henry, 1968). This can be specified in the form of restrictions on the probabilities for the LAD model in Table 1, by setting $c_1$equal to 1-b_1, e_1 equal to 1-d_1, and so on. Thus, the EISER model is more restricted than the LAD model.

The Equal True-Type-Specific Error Rates Model. Imposing the restrictions of equal error rates across the classes instead of the items, we obtain the Equal True-Type-Specific Error Rates (ETTSER) model (Clogg and Sawyer, 1981). In class 1 of Table 1 equal error rates are obtained by setting s_1 equal to a_1=b_1=d_1=f_1=h_1=j_1=m_1.. In class 2 we get the required restrictions by setting t_1 equal to a_1=b_1=d_1= f_1=h_1=j_1 (which might have different values in other classes) and 1-t_1 equal to 1-m_1. In class 3 this is done by setting u_1 equal to a_1=b_1=d_1= f_1=h_1(which values might differ in the other classes again) and 1-u_1 equal to k_1=1-m_1. Continuing this way, we arrive at class 8 which becomes the reverse of class 1.

The Proctor Model. In the Proctor model, named for the constraints suggested by Proctor (1970), all error rates – cross the classes as well as across the items – are assumed equal. In Table 1, this amounts to giving s_1 the value of $a_1 = b_1 = d_1 = f_1 = h_1 = j_1 = m_1$ (which now have equal values across the classes), and $1 - s_1$ the value of $1 - a_1 = c_1 = e_1 = g_1 = i_1 = k_1 = 1 - m_1$.

To complement the deterministic models described above we propose two mixture models, which combine a perfect Guttman pattern with measurement error, c.q. response behavior that does not totally follow the cumulative Guttman model: the ‘Guttman with guessing model’ and the ‘diagonal model’.

The ‘Guttman With Guessing’ Model. The Guttman With Guessing (GWG) model is based on two ideas. The first is Guttman’s original idea: if an item in a test is less difficult than an item already answered correctly, then this item must also be answered correctly (with probability 1). The second is that, if an item is more difficult than the person’s ability to give a correct answer, then the person guesses the correct answer.

Table 2 shows the restrictions the GWG model imposes on the conditional probabilities. To guarantee that people with a higher capability than needed by the item answer this item correctly, all conditional probabilities in the cells of the upper left part of the table must be fixed at 1. All other cells (with symbol $*$) are left free; these are estimated from the data.

The Diagonal Model. The DIAGonal model (DIAG) states that a person responds correctly with probability 1 if his ability is much higher than the item difficulty. If the ability is much lower than the item difficulty the answer is never correct (in probabilistic terms: correct with probability 0). Random behavior will occur in the turnover area from correct to incorrect. Table 3 shows how these restrictions translate into conditional probabilities given the class to which one belongs. The ideal pattern has one star only at the left-right ‘diagonal’ of the table. This model is not identified, however. We need a ‘diagonal’ of size 4 in the classes 3 to 6 to reach identification.

Table II. Guttman with guessing (the GWG model)[a]

Latent classes	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Class 1	1	1	1	1	1	1	1
Class 2	1	1	1	1	1	1	$*$
Class 3	1	1	1	1	1	$*$	$*$
Class 4	1	1	1	1	$*$	$*$	$*$
Class 5	1	1	1	$*$	$*$	$*$	$*$
Class 6	1	1	$*$	$*$	$*$	$*$	$*$
Class 7	1	$*$	$*$	$*$	$*$	$*$	$*$
Class 8	$*$	$*$	$*$	$*$	$*$	$*$	$*$

[a] The symbol $*$ means no restriction on the probability of that cell.

Table III. Random behavior at the ‘diagonal’ (DIAG model)[a]

Latent classes	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Class 1	1	1	1	1	1	$*$	$*$
Class 2	1	1	1	1	$*$	$*$	$*$
Class 3	1	1	1	$*$	$*$	$*$	$*$
Class 4	1	1	$*$	$*$	$*$	$*$	0
Class 5	1	$*$	$*$	$*$	$*$	0	0
Class 6	$*$	$*$	$*$	$*$	0	0	0
Class 7	$*$	$*$	$*$	0	0	0	0
Class 8	$*$	$*$	0	0	0	0	0

[a] The symbol $*$ means no restriction on the probability of that cell.

4. Two Simulation Studies

To classify aberrant respondents, we need latent class models that provide a good fit of nearly perfect probabilistic Guttman scale data, otherwise it remains unclear what response patterns are to be called aberrant. Therefore, we start with a simulation study to gauge how successful the various models are in fitting ‘good’ Guttman-type data. Next, we will examine how well models that fit Guttman-type data are in detecting known aberrant respondents. Models that stand both these tests are likely to detect aberrant response patterns in real data.

4.1. NORMAL DATA

In the first step, the models discussed in the previous section, with the addition of the completely deterministic original Guttman (GUT) model, are used to ana-

Table IV. Results with simulated nearly perfect Guttman data[a]

Model	Number of classes	Chi square	d.f.	Model	Number of classes	chi square	d.f.
1. LAD	8	73	108	5. GUT	8	not identified	
	9	55	100		9	225	112
	10	23	92		10	26	104
2. EISER	8	188	113	6. GWG	8	26	92
	9	48	105		9	5	84
	10	21	97		10	5	76
3. ETTSER	8	211	112	7. Diagonal	8	4	94
	9	124	104		9	2	86
	10	24	96		10	1	78
4. Proctor	8	224	119				
	9	167	111				
	10	26	103				

[a] The chi square is the log-ratio chi square. Tail probabilities are left out.

lyze the scalability of seven items that form a nearly perfect Guttman scale. The simulated data are generated from a two-parameter Birnbaum model with an item discrimination parameter of 3.4 and equally spaced item difficulties ranging from -2 to +2. This is identical to a Rasch scale model with discrimination parameter 2. We assume the latent trait θ to have a standard normal distribution in the population.

In total, 1500 cases have been generated in which 27 of the 128 (=2^7) different response patterns showed up. Only 11.5% of the 1500 generated patterns was not a perfect Guttman pattern: 9.2% showed one error (at the turnover point) and 2.3% had two or more errors. The classical reliability estimate for this data is 0.73 (Cronbach's alpha) and the scalability coefficient is 0.85 (Loevinger's H). The number of errors is small, so that each of the models of this section should be able to fit the data with 8 latent classes (for the 7 stimuli). If they cannot, we consider them too restrictive to deal with probabilistic Guttman data.[1] Table 4 shows the results for all the models by themselves, and with one, respectively two additional unrestricted error classes.

Due to the many patterns with zero expected frequency, the log-ratio chi-squares may not have a chi square distribution. Nevertheless, it seems clear that the latent distance (LAD), the Guttman with Guessing (GWG), and the Diagonal (DIAG) model are the true winners. The other models definitely need more then eight classes to fit the data well. A second criterion to select the best fitting models for these data is, whether the estimated class probabilities reproduce the relative frequencies of the Guttman patterns in the original data. Here, the GWG model and the Diagonal model perform not as well as the latent distance model, which follows the relative frequencies quite nicely. Therefore, the LAD model proves to be the best fitting model for these nearly perfect Guttman data. So, this model will be used to classify aberrant response patterns.

4.2. ABERRANT DATA

Using the LAD model, the next step is to investigate if latent class models can detect random and systematic aberrant responders. We again use data generated with the 2 parameter logistic model (Birnbaum, 1968). As before, the discrimination parameter is set at 3.4 for the seven items; the item difficulties are equidistant in the interval [−2, +2]; and the trait θ is $\sim N(0, 1)$. But now we generate 1200 'normal' cases; 150 'random aberrants' (with $\theta < 0$, and $p(X_i = 1) = 0.5$ for all items), and 150 systematic 'aberrants' ($\theta < 0$, and $p(X_i=1) =1$ for item $i = 6, 7$).

In these aberrant data 107 from the 128 possible patterns showed up; 28.1% of the 1500 cases did not have a perfect Guttman pattern, and 21.4% of the 1500 showed two or more errors. Compared with the normal data in the previous section, these data are more lifelike. This is also true for the psychometric properties of the aberrant data. Cronbach's alpha is 0.54, and Loevinger's H is 0.30; both indicate a weak scale. These data are the benchmark for the latent distance model (LAD) for Guttman-like data augmented with known aberrant respondents.

Table 5 presents the results of a number of latent class analyses with the latent distance model. As expected, we need 10 classes to fit the data adequately. In Table 5, we have two models with 10 classes. In the first of these, class 9 and class 10 are kept free. In the second, class 9 is reserved for the random aberrants and class 10 for the systematic aberrants. Random aberrants in this analysis are modeled by having the same probability of a correct answer for all items. Systematic aberrants are modeled as having random behavior at the five easiest items (their probabilities are left free), and answering correctly to the two most difficult items (with probability 1). The differences between the two models with 10 classes are small; both fit the data well. However, the model with the restricted error classes is more informative, because it provides specific information about the type of aberrance involved.

The fact that the LAD model with random and a systematic error class fits well does not guarantee that the model classifies individual respondents correctly. To investigate this, each of the 1500 simulated response patterns is assigned to the class that has the highest recruitment probability. Next, we determine the proportion normal and aberrant respondents that are classified correctly by the latent class model. These classifications are also compared with the results obtained with a classical psychometric approach, using Van der Flier's Q (1980, 1982). Q is a measure that indicates if a response pattern is aberrant. We set the significance level for the detection of aberrant response patterns using Q at alpha = 0.10. Van der Flier's Q cannot distinguish between random and systematic aberrants.[2]

Table V. Fit of the Latent distance model (LAD) for data with aberrant respondents[a]

Model	Classes		Chi square	d.f.
LAD	(8)	8 basic model	829	108
	(9)	9 + one class free	477	100
	(10)	10 +two classes free	110	92
	(11)	10 +random and systematic error	119	100

[a] The chi square is the log-ratio chi square.

Table VI. Classification errors of Van der Flier's Q and the latent distance model with the aberrant dataset

	Q	LAD
False positives:	3%	0.003%
False neg. random	Unknown	19%
False neg. system.	Unknown	12%
False neg. total	13%	16%

Using Q on the normal data classifies 168 respondents as aberrants. Since all respondents belong to the population of highly scalable respondents, the conclusion is that Q produces 11% false positives for these data. Since the LAD model fits these data successfully with 8 classes, no respondents are classified as aberrant, meaning that there are 0% false positives. The correlation of the latent trait θ with the classification scores from the latent distance model is 0.89. This is close to the correlation of θ with the sum score of the items, which is 0.91.

With the aberrant dataset, Van der Flier's Q classifies 41 of the 1200 'normal' cases as aberrant, which leads to 3% false positives in Table 6. Of the 300 aberrant patterns, 39 are classified as normal (13% false negatives). Of the random aberrants, 85% is classified correctly as aberrant by Van der Flier's method; of the systematic aberrants this is 89%. It is not possible, however, to distinguish between random and systematic aberrants.

With the latent distance model, 4 of the 1200 patterns of 'normals' are called aberrant (0.003% false positives). Of the 150 random aberrants, 29 are classified as normal and the same holds true for 18 of the 150 systematic aberrants. In Table 6 this leads to 12% and 19% false negatives respectively. Compared with

Van der Flier's Q, the model behaves quite reasonably with 16% false negatives for all aberrants, and about zero percent false positives. Furthermore, 73% of all random aberrants and 84% of all systematic aberrants are classified correctly. The correlation of the ability θ with the latent distance classification score is 0.80. The correlation of θ with the sum of the items (weighted by the LAD probability of being 'normal') is even higher: 0.85. Taking in mind that the correlation with the unweighted sum is 0.75, we may conclude that the latent distance model with extra classes for cheaters and guessers adds information by modeling the errors.[3]

5. Discussion

Various models have been presented as possible models for response behavior on a cumulative scale. The descriptions are in terms of cognitive items that can be right or wrong. However, the models have an interpretation in terms of attitudinal items as well. This is most clear in the probabilistic LAD model, and the models derived from the LAD. There, persons differ in attitude and items differ in extremeness. The response process follows this structure: extreme items are more difficult to agree with, and only persons with an extreme attitude will agree to them. The mixed models, the Guttman with guessing (GWG) and the diagonal (DIAG) model, are interesting, because they embody a different theoretical response process. In the GWS model, the response process for cognitive items is guessing when the items become too difficult. Translated to attitude items, as long as we are positive about a specific item we say yes with probability 1. However, if the items become more extreme than our own point of view, we are going to hesitate and show a tendency for random behavior. The DIAG model embodies another response process, which specifies that persons have a deterministic response process when the items are either very easy or very difficult, relative to their own position. Compared with the GWG model this model has a clear interpretation for ordered attitudinal items. If we are definitely positive about a specific item, we say 'yes' with probability 1 to this item, and to the more 'easy' items. If we are unmistakably against, we do not agree and say 'yes' with probability 0 to this item and to the more 'difficult' items. For items between these two extremes we are hesitating and say sometimes 'yes', and sometimes 'no'. We show more or less random behavior in that case. Since people do not guess at the more difficult items in this model, the interpretation for cognitive items is more intricate. We might think of catch-questions in which people with ability lower than the item difficulty are seduced into giving the wrong answer.

If the data contain many perfect Guttman patterns when the items are ordered to 'difficulty', and if the log-ratio chi square indicates that a model fits well, the LCA results still need not be what might be expected according to Guttman's ideas. The LCA models presented here do not impose order restrictions on the estimated probabilities, as is done in, for instance, the models proposed by Croon (1991). Thus, the result can be that the order does not follow the Guttman model that underlies these LCA models. In our simulation study with the normal data, that follow a Rasch model, only the LAD and GWG models performed reasonably well in this respect. The other models failed to reproduce the observed probabilities. This may be the result of specific characteristics of the data, but a clear explanation fails. The LAD model finally chosen links up properly with the remaining models, and it is plausible with respect to the frequency distribution of the perfect Guttman patterns in the original data. In an empirical study (Van den Wittenboer et al., 1997), we found that the LAD model performed well in diagnosing aberrance in a sample of elderly respondents.

The discussion above focuses on interpretation and response processes. A technical problem encountered with all models, whether applied to simulated or real data, is the sensitivity of the latent class solution to starting values. Both in LCAG (Hagenaars and Luijkx, 1987) and Panmark (Van der Pol et al., 1991) the estimation procedure leads to a local maximum most of the time. Panmark has the option to generate a large set of random starting values and to compare the results of the (default) first eight iterations. This default value was not enough, however. Even after 1000 starting values, we could get into local maxima.[4] Only after using twenty (time consuming) initial iterations for a thousand random starting values, to select the most promising set of starting values for the estimation procedure, the results became trustworthy. The problem, however, is that most analysts will expect a program to produce maximum likelihood estimates, and may not be aware of the problem of encountering a local maximum. Especially for the inexperienced user, this is a serious drawback of latent class analysis with computer programs using the EM algorithm. It would be nice, if a statistical method leading to starting values could be used that in some sense guarantees that a real maximum will be found.

A related technical problem is that the computer programs show a strong tendency to converge on probabilities equal to zero or one. Guttman classes with relatively low frequencies, for example, which are nevertheless present in the data, are often set equal to zero instead of a small probability as one should expect. It is not obvious whether this is because of the data, the specific models, or the EM algorithm used in the analysis. One problem that is related to the algorithm is that, once a latent class is estimated to have probability zero, this value cannot change in subsequent iterations.

To deal with perfect Guttman classes with ‘zero’ probability, Clogg and Sawyer (1981) suggest that the item which causes the problem can be eliminated under certain conditions. It remains unclear, however, which item this should be. Suppose, for example, that we have 4 items ordered from A to D which form a perfect Guttman scale. Suppose furthermore that the class with pattern 1112 has a probability of zero. If we now look at the four remaining classes (1111), (1122), (1222), and (2222), we only observe that the pattern of responses across the classes is exactly the same for item C as it is for item D. But is this because of item D, or of item C? In a perfect Guttman scale such a question would not matter; we can leave out C or D without losing information. In a probabilistic environment where the errors also give information about the scalability, it is unclear what should be done. Therefore, we decided to leave the scale intact. Another suggestion might have been leaving out the item that causes most of the errors in the otherwise perfect Guttman patterns, so that we optimize the number of correct Guttman patterns. For the time being this should be done by hand, while it is still not clear whether this is the best solution.

Even if none of these problems did interfere with the selection of the final model, there remains the definition of the two types of errors in the extra latent classes. Aside from the fact that there are more aberrant response patterns than these two, albeit more difficult to formulate as restrictions for the computer programs, the definition of the random aberrants is perhaps to restrictive for random behavior. In fact, it is equal to the first class of the too restrictive 'equal true-type-specific error rates' model, which also gives rise to equal probabilities of positive answers to the items. Perhaps, this might be the reason that low probabilities show up for this class. On the other hand, if this class is left free completely, the information is too unsubstantial to draw a conclusion about the kind of errors people that made. Actually, we ought to have some set of restrictions between completely free and equal response probabilities, preferably based on theory.

It is interesting to note in this context that systematic aberrance as we defined it contains restrictions of this type. It allows free probabilities for some items, while the response probabilities for the most difficult items are fixed at one. Unfortunately, systematic aberrants of this type are scarce in our real life data. Relaxing the restrictions to a probability of one for errors in the most difficult item of this class only would interfere too much with the usual error probability that should be allowed for probabilistic Guttman scales. Nevertheless, we are convinced that error modeling by means of restrictions in latent class analysis leads to a better understanding of aberrant response behavior.

Notes

1. It is impossible to improve the likelihood by using more than $(j+1)/2=4$ latent classes if we analyze the data as a mixture Rasch model (De Leeuw and Verhelst, 1986; Lindsay et al., 1991). Since we analyze the data with highly restricted conventional latent class models, this rule.does not apply here. Even with the two extra classes for intrinsically unscalable respondents (cf. Goodman, 1975) the results in Table 4 show a strong decrease of the chi-square.
2. Van der Flier's Q, more completely denoted by $Q(x)$ is the right-tail probability of a response pattern within the conditional distribution of pattern probabilities, given the probabilities of the questions of the scale. Thus, $Q(x)$can be interpreted as the one-tailed significance level (p) used in statistical testing. That is, a small value of $Q(x)$ indicates that the probability to find this specific response pattern is small. Therefore, that pattern is unexpected or aberrant. A large value means that a respondent has a response pattern on a set of questions as could be expected. The assumptions made for the calculation of $Q(x)$ are identical to the assumptions of Mokken's model of monotone homogeneity (monotonicity in the latent trait, local stochastic independence, and unidimensionality).

3. We also examined the results with the ‘Guttman with Guessing’ (GWS) model. This proved somewhat inferior to the LAD model, both in its capacity to correctly detect aberrant response patterns and in producing estimates of the latent trait.
4. A local maximum is suspected when a restricted model has a smaller chi-square than a less restricted model, or when the parameter estimates behave erratically across similar models.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In: F. M. Lord and M. R. Novick (eds), *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Clogg, C. C. (1988). Latent class models for measurement. In: R. Langeheine & J. Rost (eds), *Latent Trait and Latent Class Models.* New York: Plenum Press.

Clogg, C. C. & Sawyer, D. O. (1981). A comparison of alternative models for analyzing the scalability of response patterns. In: S. Leinhardt (ed.). *Sociological Methodology 1981.* San Francisco: Jossey-Bass, pp. 240–280.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology* 44: 315–331.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.

Van der Flier, H., (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-cultural Psychology* 13: 267–298.

Formann, A. K. (1988). Latent class models for nonmonotone dichotomous items. *Psychometrika* 53: 45–62.

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

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

Groves, R. M. (1989). *Survey Errors and Survey Costs.* New York: Wiley.

Hagenaars, J. A. & Luijkx, R. (1987). *LCAG; Latent Class Models and other Log-linear Models with Latent Variables.* User’s manual, Working Paper Series #17, Tilburg University, Department of Sociology.

Harnish, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement* 18: 133–146.

Heinen, A. G. J. J. (1993). *Discrete Latent Variable Models.* Tilburg: Tilburg University Press.

Langeheine. R. (1988). New developments in latent class theory. In: R. Langeheine & J. Rost (eds), *Latent Trait and Latent Class Models.* New York: Plenum Press.

Lazarsfeld, P. F. (1950). The interpretation and computation of some latent structures. In: S. A. Stouffer, L. Guttman, E. Suchman, P. F. Lazarsfeld & J. Clausen (eds), *Measurement and Prediction.* Princeton: Princeton, University Press, pp. 413–472.

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

De Leeuw, E. D. & Hox, J. J. (1994). Are inconsistent respondents consistently inconsistent? A study of several nonparametric person fit indices. In: J. J. Hox & W. Jansen (eds), *Measurement Problems in the Social Sciences*. Amsterdam: SISWO, pp. 67–88.

De Leeuw, J. & Verhelst, N., (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics* 11: 183–196.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test score. *Journal of Educational Statistics* 4: 269–290.

Lindsay, B., Clogg, C. C. & Grego, J., (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* 86: 96–107.

McCutcheon, A. L. (1987). *Latent Class Analysis.* Newbury Park: Sage.

Meijer, R. R. (1994). Nonparametric person fit analysis. PhD Thesis, Amsterdam: Vrije Universiteit.

Meijer, R. R. & De Leeuw, E. D. (1993). Person fit in survey research: The detection of respondents with unexpected response patterns. In J. H. L. Oud & R. A. W. van Blokland-Vogelesang (eds), *Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences*. Nijmegen: ITS, pp. 236–245.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis.* The Hague, The Netherlands: Mouton.

Molenaar, I. W. & Hoytink, H. (1990). The many null distributions of person fit indices. *Psychometrika* 55: 75–106.

Van der Pol, F., Langeheine, R. & De Jong, W. (1991). *Panmark User Manual; Panel Analysis Using Markov Chains*. Version 2.2. Voorburg, CBS, Department of Statitical Methods.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika* 35: 73–78.

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

Rost, J. & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (eds), *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Münster/New York: Waxmann, pp. 13–37.

Sijtsma, K. (1988). *Contributions to Mokken's Nonparametric Item Response Theory*. Amsterdam, The Netherlands: Free University Press.

Tarnai, C. & Rost, J. (1990). *Identifying Aberrant Response Patterns in the Rasch Model*. Munster: Sozialwissenschaftliche Forschungsdokumentationen.

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns. *Journal of Educational Statistics* 7: 215–231.

Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement* 20: 221–230.

Van Tilburg, T. G. & De Leeuw, E. D. (1991). Stability of scale quality under various data collection procedures: A mode comparison on the "De Jong-Gierveld loneliness scale". *International Journal of Public Opinion Research* 3: 69–85.

Van den Wittenboer, G., Hox, J. J. & De Leeuw, E. D. (1997). Aberrant response patterns in elderly respondents: Latent class analysis of respondent scalability. In J. Rost & R. Langeheine (eds), *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Münster/New York: Waxmann, pp. 15–162.