# Statistical models for categorical variables

van der Ark, L.A.; Croon, M.A.; Sijtsma, K.

*Published in:*
New developments in categorical data analysis for the social and behavioral sciences.

*Publication date:*
2005

*Citation for published version (APA):*
van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2005). Statistical models for categorical variables. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences.* (pp. 1-14). Lawrence Erlbaum.

# Chapter 1

# Statistical Models for Categorical Variables

L. Andries van der Ark, Marcel A. Croon, and Klaas Sijtsma
*Tilburg University*

This volume contains a collection of papers on the analysis of categorical data by means of advanced statistical methods. Most methods presented use one or more latent variables to explain the relationships among the observed categorical variables. If the latent variables are also categorical the method is called latent class analysis (LCA) and if they are continuous the method is called item response theory (IRT) or latent variable modelling.

Both LCA and IRT are used to analyze categorical data from at least two, but often many variables collected in a multidimensional contingency table. It is for this reason that this introductory chapter starts with a brief introduction into the analysis of contingency tables, and then introduces log-linear models, LCA and IRT at a conceptual level. The chapter ends with a brief outline of the contributions to this volume.

The focus of the contributions is applied; that is, after a method is explained, the potential of the method for analyzing categorical data is illustrated by means of a real data example. The editors express their hope that this volume is helpful in guiding applied researchers in the social and the behavioral sciences, and possibly in other fields (e.g., language studies, marketing, political science, social medical research) as well, to use the advanced and multi-purpose models discussed here in their own research.

## 1.1 Categorical Data and Analysis of Contingency Tables

Many variables collected in social and behavioral science research are categorical. Agresti (2002) distinguishes two kinds of categorical variables. Nominal variables have two or more numerical values that distinguish classes, for example, gender [men (e.g., score 0) and women (e.g., score 1)], religion [e.g., catholic (1), protestant (2), jewish (3), islamic (4)], and political persuasion [democratic (1), republican (2)]. The scores serve to distinguish group membership. Ordinal variables have numerical values that describe an ordering. Examples are level of education [low (1), intermediate (2), high (3)], preference for a brand of beer (e.g., scores $1, \ldots, 10$; a higher score indicates a stronger preference), and level of agreement with a particular statement about abortion (e.g., $0, \ldots, 4$; a higher score indicates a stronger endorsement). In these examples, the scores serve to order the respondents on the variable of interest.

Traditionally, relationships between categorical variables are studied by means of contingency tables. The simplest contingency table gives the two-dimensional layout for two variables, such as gender and political persuasion. In the example, the table has two rows (gender) and two columns (political persuasion). For a sample of respondents, the cells of the table give the number of democratic men, republican men, democratic women, and republican women. The margins of the table give the sample frequency distributions of gender and political persuasion. An example of such a simple two-way contingency table is Panel A of Table 2.6 on page 29.

The relationship between gender and political persuasion can be studied by means of several statistics. For tables of any dimension (i.e., number of variables) and any order (i.e., number of categories per variable), the chi-square statistic (denoted $\chi^2$) can be used to test a null hypothesis of expected cell frequencies under a particular model for the data against an unspecified alternative. An example of a more general two-way contingency table is Table 5.1 on page 85. These expected frequencies can only be calculated under certain assumptions about the population. A common assumption is that the observed marginal distributions are the population distributions, which are kept fixed, and used to calculate cell frequencies expected under independence of the variables. The assumption then is that the variables' marginal distributions generated the data. If the chi-squared statistic is significant, the null hypothesis is rejected and it is inferred that there is a relationship between the variables. Other null models can be formulated, and expected cell frequencies calculated and tested against the observed cell frequencies, using the chi-squared statistic.

For some simple tables, the strength of the relationship can be expressed by association coefficients. For example, in a $2 \times 2$ table for gender (rows; scored 0, 1) and political persuasion (columns; scored 1, 2), assuming the marginal distributions fixed one can easily verify that the table has one degree of freedom. Consider the "democratic men" cell [i.e., the (0, 1) cell]. Given fixed marginals and a given sample size, $N$, the expected frequency of democratic men can be calculated and compared with the observed frequency. Obviously, the observed frequency can deviate from the expected frequency by being either higher or lower, and the more it deviates in either direction, the stronger the relationship. The strength of this relationship can be expressed by the $\phi$ coefficient. To define this coefficient, denote the cell frequencies of the gender by political persuasion table as $n_{01}$, $n_{02}$, $n_{11}$, and $n_{12}$, and the marginal frequencies of the rows as $n_{0+}$ and $n_{1+}$, and of the columns as $n_{+1}$, and $n_{+2}$. The $\phi$ coefficient is defined as

$$\phi = \frac{n_{01}n_{12} - n_{02}n_{11}}{\sqrt{n_{0+}n_{1+}n_{+1}n_{+2}}}.$$

The dependence of $\phi$ on the $\chi^2$ statistic is clear through

$$\phi = \sqrt{\frac{\chi^2}{N}}.$$

This relationship shows that, given fixed $N$, the higher $\chi^2$ (i.e., the greater the discrepancy between observed and expected cell frequencies), the higher $\phi$, either positive or negative (i.e., the stronger the association between gender and political persuasion). The $\phi$ coefficient is equal to the Pearson product-moment correlation, applied to the respondents' nominal scores on gender (0,1) and political persuasion (1,2), and thus attains values on the interval $[-1; 1]$ provided that the marginal distributions of the variables are equal. The more the marginal distributions are different, the smaller the range of $\phi$.

The dependence of $\phi$ on the marginal distributions of the table obviously impairs its interpretation. This drawback is remedied by Mokken's (1971; see Loevinger, 1948) $H$ coefficient. Let $\phi_{max}$ denote the maximum correlation given the marginals of the table; then, the $H$ coefficient is defined as

$$H = \frac{\phi}{\phi_{max}}.$$

Division by $\phi_{max}$ guarantees that the maximum $H$ is always 1.

Probably the best known association coefficient for contingency tables is the odds-ratio (e.g., Agresti, 2002, pp. 44-47). For cell frequencies denoted

$n_{01}$, $n_{02}$, $n_{11}$, and $n_{12}$, the odds ratio, denoted $O$, is defined as

$$O = \frac{n_{01}n_{12}}{n_{02}n_{11}}.$$

It takes values in the interval $[0, \infty)$. An odds ratio smaller than 1 indicates a negative relationship and an odds ratio greater than 1 a positive relationship. The odds ratio is not influenced by the marginal distributions of the table. All three coefficients can be generalized to two-way contingency tables of greater order.

Another example of a contingency table in which association can be determined is the following. Imagine two psychologists who independently rated children's inclination to engage in self-directed behavior. Assume that inclination is taken as the degree to which children exhibit this kind of behavior, recorded by means of a checklist, when they are observed in a playground among their peers. Assume that both psychologists observe each child for a fixed period of time, and then rate the child as either "low-level," "average," or "high-level." For $N$ rated children, the ratings of the two psychologists can be collected in a $3 \times 3$ contingency table, with diagonal cells containing the frequencies by which they agreed, and the off-diagonal cells the frequencies by which they disagreed. The marginal distributions express each psychologist's propensity for assigning children from the population of interest to the three categories. Assuming the marginals fixed, the expected frequencies can be calculated and compared to the observed frequencies, using a chi-squared test. The degree to which the psychologists agree can be expressed by means of Cohen's $\kappa$ coefficient, which is normalized to have a maximum of 1 independent of the marginals, expressing maximum agreement, a 0 value expressing independence, and a negative minimum which depends on the marginals. Cohen's $\kappa$ has been generalized to more raters and different numbers of categories used per rater, and also the differential weighing of the off-diagonal cells.

The methods discussed thus far are suited especially for small tables, but may miss several interesting effects in tables based on more variables and/or more categories per variable. For example, consider a $3 \times 4$ table with level of education in the rows (low, intermediate, high) and religion in the columns (catholic, protestant, jewish, islamic). A researcher may hypothesize, in particular, that people with a low educational level are more often catholic than expected on the basis of the marginal frequencies of low educational level and catholic religion alone. Similarly, he/she may expect higher or lower frequencies elsewhere in the table, but not everywhere. The overall chi-squared statistic and the association coefficients cannot reveal such specific effects, but log-linear models can.

Conceptually, log-linear models may be compared with analysis of variance models (Stevens, 1992, p. 502). Log-linear models compare effects of rows and columns of a contingency table with a "grand mean" and are also capable of explaining deviates from marginal effects in cells by means of interaction effects. Let $X$ and $Y$ be two categorical variables with values $x = 1, \ldots, m_X$ and $y = 1, \ldots, m_Y$, respectively. The natural logarithm of the expected frequency in cell $(x, y)$, denoted $e_{xy}$, is modelled to be the sum of a grand mean, $\lambda$, a row effect, $\lambda_x^X$, a column effect, $\lambda_y^Y$, and an interaction effect, $\lambda_{xy}^{XY}$, such that

$$\ln e_{xy} = \lambda + \lambda_x^X + \lambda_y^Y + \lambda_{xy}^{XY} .$$

Log-linear models that contain all main effect and interaction effect parameters—so-called saturated models—cannot be tested because there are no degrees of freedom left in the data. More importantly, the principle of parsimony requires models to be as simple as possible and this is realized best when the researcher defines the effects of interest before he/she starts analyzing the data. The other effects can be set to 0, comparable to what one does with the factor loadings in a confirmatory factor analysis, and the fit of the restricted model to the data can be tested using chi-squared test statistics. Also, competing models which are nested can be tested against one another. For example, for cell $(x, y)$ nested models with and without the interaction parameter can be tested against each other. A significant result means that observed frequency is different from the expected frequency under the null model. See Wickens (1989), Hagenaars (1990), Agresti (1996, 2002), Stevens (1992), and Andreß, Hagenaars, and Kühnel (1997) for more information on log-linear models; also see Bergsma and Croon (chap. 5, this volume).

## 1.2 Categorical Data and Latent Class Analysis

LCA models assume that the frequency counts in a contingency table can be explained by finding an appropriate subgrouping of respondents, such that in each table corresponding to a subgroup the cell frequencies can be explained from the marginal distributions for that table. As the subgroups are not defined a priori but estimated from the data, they are considered to be latent; hence, latent class analysis.

Assume a discrete latent variable on which homogeneous classes of respondents can be distinguished, and denote this variable $\theta$, with $W$ classes, indexed $w = 1, \ldots, W$. Also, assume an arbitrary number, say, $J$, of observed categorical variables, denoted $X_j$, with $j = 1, \ldots, J$, and collected in a vector $\mathbf{X}$ with realization $\mathbf{x}$. The LCA model assumes independence between the observed variables given a fixed value of $\theta$. This is known as local independence (LI), which means that

$$P(\mathbf{X} = \mathbf{x} \mid \theta = w) = \prod_{j=1}^{J} P(X_j = x_j|\theta = w).$$

Now, using the property of independent events, $A$ and $B$, that $P(A \wedge B) = P(B)P(A)$, and applying LI to $P(A)$, we may write the LCA model as (Goodman, 2002; Heinen, 1996, p. 44; McCutcheon, 2002),

$$P(\mathbf{X} = \mathbf{x} \wedge \theta = w) = P(\theta = w) \prod_{j=1}^{J} P(X_j = x_j|\theta = w).$$

The probability that a randomly chosen respondent produces score pattern $\mathbf{X} = \mathbf{x}$, is

$$P(\mathbf{X} = \mathbf{x}) = \sum_{w=1}^{W} P(\theta = w) \prod_{j=1}^{J} P(X_j = x_j|\theta = w).$$

This equation shows how the LCA models the $J$-variate distribution of the observable variables in terms of latent class probabilities, $P(\theta = w)$, and probabilities of having particular scores $x_j$ on observable variable $X_j$ ($j = 1, \ldots, J$) given class membership, $P(X_j = x_j|\theta = w)$.

The class probabilities and the conditional probabilities can be estimated from the data for several choices of the number of latent classes, $W$. In practical data analysis, $W$ often varies between 1 and 5. The parameter estimates for the best-fitting model are used to estimate the discrete distribution of $\theta$, $P(\theta = w)$, with $w = 1, \ldots, W$. This distribution can be used together with the conditional probabilities, $P(X_j = x_j|\theta = w)$, to assign people to latent classes. For respondent $v$, this is done using probabilities $P(\theta = w|\mathbf{X}_v = \mathbf{x}_v)$, for $w = 1, \ldots, W$, after which he/she is assigned to the class that has the greatest subjective probability.

For a given number of latent classes, one thus finds a typology for a population in terms of response patterns on $J$ variables; that is, different classes are characterized by different patterns of scores on the $J$ observable variables. For example, a sociologist may be interested in the types of attitudes with respect to male and female role patterns, interpreted in terms of the typical answer pattern on the $J$ items in a questionnaire. Heinen (1996, pp. 44-49) found that three classes fitted the data best. One class (45% of the respondents) represented a pro-women's lib point of view, another class

(11%) was traditional, and the third (44%) was liberal on some issues but traditional on others. Another example comes from developmental psychology, where researchers may be interested in different developmental groups. Each group may be characterized by another solution strategy for a particular cognitive problem, which reflects the cognitive stage of the group (e.g., Bouwmeester, Sijtsma, & Vermunt, 2004; Jansen & Van der Maas, 1997; Laudy, Boom, & Hoijtink, chap. 4, this volume).

It may be noted that, thus far, latent classes have been assumed to be unordered, that is, to have nominal measurement level. This leads to an unrestricted LCA model. A recent development is to put order restrictions on the conditional probabilities, $P(X_j = x_j|\theta = w)$, so as to express the assumption that there is an ordering among the latent classes, such that people in a higher latent class have a higher probability, $P(X_j = x_j|\theta = w)$, to give a particular answer to the item. This makes sense, for example, when a higher class stands for a higher reading ability and the items contain questions about a reading text, or when higher classes correspond to progressively higher levels of endorsement with abortion and the items are positively worded statements about abortion that have to be answered on a rating scale. Croon (1990, 2002) introduced these ordered latent class models, which were studied further by Hoijtink and Molenaar (1997), Vermunt (2001) Vermunt and Magidson (chap. 3, this volume), and Van Onna (2002); see Emons, Glas, Meijer, and Sijtsma (2003) for an application to the analysis of odd response patterns on sets of cognitive test items, and Laudy, Boom, and Hoijtink (chap. 4, this volume) for a application to balance-scale data. Haberman (1979) and Heinen (1996) discussed the close mathematical relationships between log-linear models and LCA models. More information on LCA models can be found in McCutcheon (1987) and Hagenaars and McCutcheon (2002).

## 1.3 Categorical Data and Item Response Theory

IRT models assume that the frequency counts in a $J$-dimensional contingency table based on the $J$ items from a test can be explained by one ore more continuous latent variables on which the respondents are located. For one latent variable, given a fixed value, each contingency table corresponding to this value can be explained from the marginal distributions for that table. This is the assumption of local independence (LI) for IRT models.

The item scores usually are ordinal, expressing progressively higher levels of endorsement (e.g., Masters, 1982; Samejima, 1969), and sometimes nominal, as with multiple-choice items when students select one option from four or five unordered options (e.g., Bock, 1972; Thissen & Steinberg, 1997). Assume that we have $Q$ continuous latent variables, enumerated $\theta_1, \ldots, \theta_Q$, and collected in vector $\boldsymbol{\theta}$. Then, LI is defined as

$$P(\mathrm{X} = \mathrm{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} P(X_j = x_j | \boldsymbol{\theta}).$$

For simplicity we assume that one latent variable, $\theta$, suffices to explain the data structure, and that the probability density of $\theta$ is denoted $g(\theta)$. Then, the multivariate distribution of the data can be written as

$$P(\mathrm{X} = \mathrm{x}) = \int_{\theta} \prod_{j=1}^{J} P(X_j = x_j | \theta) g(\theta) d\theta.$$

The difference with the multivariate distribution of the data in an LCA model is that in IRT the latent variable is continuous, thus introducing an integral instead of a summation, while in LCA models the latent variable is discrete.

IRT models impose restrictions on the conditional response probabilities, $P(X_j = x_j|\theta)$. These restrictions can be orderings only (e.g., the response probability increases in the latent variable) or consist of the choice of a parametric function, such as the normal-ogive or the logistic. Once a model is chosen, it is fitted to the data. If a misfit is obtained, either the restrictions on the conditional response probabilities, $P(X_j = x_j|\theta)$, or the dimensionality of the model are changed, or items that were badly fitted by the model are removed from the analysis. Either way, the new model is fitted to the complete data set or the original model is fitted to the modified data set. When a fitted model is obtained, parameter estimates for items are used to calibrate a scale ($\theta$) for respondents, on which respondent $v$ is located by means of ML (or Bayesian) estimates of $\theta_v$ ($v = 1, \ldots, N$). This scale is then used as a measurement rod for the psychological property operationalized by the items.

Many applications of IRT models exist. For example, they are used to build large item collections—item pools—with known measurement properties (Kolen & Brennan, 1995), from which tests with desirable properties can be assembled (Van der Linden, 1998). Item pools are also the basis of computerized adaptive testing, which is the one by one adminstration of items to individuals where the choice of the next item is determined by the estimate of the individual's value on the latent variable based on the previous items, until an estimate of sufficient accuracy is obtained (Van der Linden & Glas, 2000). IRT models are also used to detect items that are biased against a particular minority group (Holland & Wainer, 1993), and individuals that show atypical test performance (Meijer & Sijtsma, 2001). Another application is the study of the cognitive process underlying the item responses by means of an appropriate re-parametrization of the item parameters in an IRT model (e.g., Fischer, 1974; Embretson, 1997).

An interesting development in the 1990s has been that IRT models have become part of a larger, more encompassing statistical tool box. For example, they have become the measurement part of linear hierarchical models for analyzing nested data (Fox, chap. 12, this volume; Fox & Glas, 2001; Patz, Junker, Johnson, & Mariano, 2002). This development in IRT is comparable to the integration of multilevel models, event-history models, regression models, and factor analysis models in LCA. See Vermunt (1997) for the development of the general framework in which these models were incorporated. Both developments reflect the increased availability of advanced statistical machinery for analyzing complex data, integrating structural analysis with the analysis of differences between groups (LCA) or individual differences (IRT).

Another interesting development is the integration of LCA and IRT. For example, nonparametric IRT models often restrict the conditional probabilities, $P(X_j = x_j|\theta)$, to be nondecreasing. This assumption reflects the idea that a higher latent variable value, for example, arithmetic ability increases the probability of solving arithmetic problems correctly. This monotonicity assumption has recently inspired the approximation of continuous nonparametric IRT models by discrete ordered LCA models (Hoijtink & Molenaar, 1997; Van Onna, 2002). The idea is that a small number of ordered latent classes can approximate the continuous latent variable with sufficient accuracy, and then make available for nonparametric IRT the repertoire of standard statistical techniques needed for investigating model fit. LCA models have also been used in the context of parametric IRT models, for example, the Rasch model; see Rost (1990). Introductions to IRT models can be found in Embretson and Reise (2000), Fischer and Molenaar (1995), Van der Linden and Hambleton (1997), and Sijtsma and Molenaar (2002).

## 1.4 Contents of This Volume

Hagenaars (chap. 2) discusses how misclassification and measurement errors in categorical variables lead to phenomena that are similar to the well-known regression toward the mean effect for continuous variables. He argues that for categorical variables one should rather speak of tendency toward the mode and shows by means of well-chosen examples how frequently this phenomenon occurs in social science research. He also discusses how tendency toward the mode can be fixed by appropriate latent class analyses of the data.

Vermunt and Magidson (chap. 3) attempt to bridge the differences between the linear factor analysis model for continuous data and the latent class model for categorical data. In their approach, a linear approximation to the parameter estimates obtained under a particular latent class model, the latent class factor analysis model is obtained. By means of this model they ensure that the output of their analysis is similar to that of standard factor analysis, which may be easier to interpret than the output from the original LCA.

Laudy, Boom, and Hoijtink (chap. 4) use LCA to test hypotheses involving inequality restrictions (e.g., Group A is expected to perform better on test T than Group B), and discuss how a researcher may choose among competing hypotheses. The authors analyze categorical balance-task data obtained from 900 children of different age groups, and compare several theories explaining the associations in these data.

Bergsma and Croon (chap. 5) discuss a broad class of models for testing complex hypotheses about marginal distributions for categorical data. The models are defined by means of the nonlinear equality constraints imposed on the cell probabilities in the corresponding contingency table. Furthermore, the authors discuss how the maximum likelihood estimates of these constrained cell probabilities may be obtained, and how the corresponding model can be tested.

Moustaki and Knott (chap. 6) use the EM estimation procedure and a Bayesian estimation procedure to estimate the parameters of three latent variable models for categorical data. The authors demonstrate how latent variable models for categorical data can be formulated on the basis of substantial theory. They discuss the merits and the pitfalls of both estimation procedures using software that is freely available.

Van Rijn and Molenaar (chap. 7) discuss dynamic latent variable models that allow the analysis of categorical time series observations on a single subject. The authors describe a model that integrates the basic principles of the Rasch measurement model and the assumptions of a simple stochastic model for describing individual change. They also discuss parameter estimation for this dynamic Rasch model.

Van der Ark and Sijtsma (chap. 8) discuss the imputation of item scores for missing values in data stemming from the administration of tests and questionnaires. They consider simple and more complex methods for missing data handling and single and multiple imputation. They apply their methods to three real data sets in which a priori fixed numbers of scores have been deleted artificially using several missing data mechanisms, and then impute scores for the missingness thus created. The effects of each imputation method on confirmatory and exploratory IRT scale analysis is investigated.

Kelderman (chap. 9) formulates measurement models for categorical data in terms of graphical independence models, exchangeability models, and log-linear models. By bringing these concepts under a single umbrella, he demonstrates how to start from scratch and construct an IRT model using important concepts such as exchangeability and internal and external consistency as building blocks.

Bechger, Maris, Verstralen, and Verhelst (chap. 10) discuss the Nedelsky IRT model for the analysis of test data from multiple-choice items on which some of the examinees may have guessed for the correct answer. The model rests on the assumptions that an examinee first eliminates the item's distracters he or she recognizes to be incorrect, and then guesses at random from the remaining options. They apply the model to data from a national test administered to eighth grade elementary school pupils, assessing their follow-up school level.

Draney and Wilson (chap. 11) discuss the saltus IRT model. This is a mixture Rasch model. The saltus model is especially suited for developmental data stemming from several subpopulations who are in different developmental stages. The model assumes one item difficulty parameter for each item and formalizes change from one subpopulation to the next by means of a small number of parameters. The authors apply the saltus model to data obtained from 460 children ranging in age from 5 to 17 years, who took a test assessing proportional reasoning, and show how the data should be analyzed and the results interpreted.

Fox (chap. 12) discusses a multilevel IRT model. He analyzes the data of a mathematics test administered to 2196 pupils (level 1 of the multilevel IRT model) from 97 elementary schools (level 2). This chapter focuses on the goodness of fit of the multilevel IRT model and the detection of outlying response patterns.

## References

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Andreß, H. J., Hagenaars, J. A. P., & Kühnel, S. (1997). *Analyse von Tabellen und kategorialen Daten* (Analysis of tables and categorical data). Berlin: Springer.

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

Bouwmeester, S., Sijtsma, K., & Vermunt, J. K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology, 1*, 67-86.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171-192.

Croon, M. A. (2002). Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 137-162). Cambridge: Cambridge University Press.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-321). New York: Springer.

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

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*, 459-478.

Fischer, G. H. (1974). *Einfürung in die Theorie psychologischer Tests* (Introduction to psychological test theory). Bern, Switzerland: Huber.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models. Foundations, recent developments, and applications.* New York: Springer.

Fox, J. -P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271-288.

Goodman, L. A. (2002). Latent class analysis. The empirical study of latent types, latent variables, and latent structures. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 3-55). Cambridge: Cambridge University Press.

Haberman, S. J. (1979). *Analysis of qualitative data*, 2 vols. New York: Academic Press.

Hagenaars, J. A. P. (1990). *Categorical longitudinal data. Log-linear, panel, trend, and cohort analysis.* Newbury Park, CA: Sage.

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

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-189.

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

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

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

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin, 45*, 507-530.

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

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, CA: Sage.

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin: De Gruyter.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341-384.

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

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

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

Stevens, J. (1992). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

Van der Linden, W. J. (1998). Special issue of *Applied Psychological Measurement* on "Optimal test assembly," *22*, 195-302.

Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing. Theory and practice.* Dordrecht, The Netherlands: Kluwer.

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

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519-538.

Vermunt, J. K. (1997). *Log-linear models for event histories.* Thousand Oaks, CA: Sage.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283-294.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum.