Psychologica Belgica
1998, 38-2, 61-82.

INVITED ADDRESS

MODELING LOCAL ITEM DEPENDENCIES IN ITEM RESPONSE THEORY

Francis TUERLINCKX and Paul DE BOECK
University of Leuven

In this paper, an introduction to some existing Item Response Theory (IRT) models for Local Item Dependencies (LIDs) is presented together with an application. First, four LID models introduced by Hoskens and De Boeck (1997) are discussed and a more general model is derived. Second, it is described in broad outline how these models can be estimated and tested statistically via a loglinear approach. Third, responses are examined on questions that cannot be understood by the respondents. For this purpose a Polish Questionnaire was used in which there was no clear indication in the items of which response (yes or no) had to be chosen. An interpretable dependency structure was found for the first part of the test, but it could not be validated for the entire test. In conclusion, the problems are discussed that are inherent to the classical estimation and testing procedure when dealing with local item dependencies and some solutions are offered.

Theories in psychology and social sciences often use unobservable concepts like intelligence, social class or depression. Those concepts, called latent variables, are assumed to form the underlying structure behind observable phenomena; covariations between observations are attributed to these latent traits. The observable phenomena, called manifest variables, are used to infer information about the latent variables. In this paper we consider only continuous latent variables, called latent traits, and discrete manifest variables, like items or questions with two or more response categories. The relation between the latent traits and the responses to items is often stated in a mathematical model. Item response theory (previously known as latent trait theory) is a collective term for mathematical models for the relation between latent characteristics of persons and items on the one hand and the response of these persons on the items on the other hand. An introduction to IRT models can be found in Fischer

---

In 1997, Francis Tuerlinckx won the Award for the Best License Thesis of the Belgian Psychological Society. The present Invited Address is a part of the award. It presents more recent work.

The first author is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium). He wishes to thank the Belgian Psychological Society for being awarded with the "Best Thesis Award" (1997), Willy Lens for the possibility to collect the data and Tom Verguts and two anonymous reviewers for helpful comments.

Correspondence concerning this article should be addressed to Francis Tuerlinckx, Tiensestraat 102, B-3000 Leuven, Belgium. Electronic mail may be sent to Francis.Tuerlinckx@psy.kuleuven.ac.be.

and Molenaar (1995), Van der Linden and Hambleton (1997) and Verhelst (1993).

A very simple IRT model is the Rasch model (Rasch, 1980). Suppose a person S_v is presented an item I_i, which has two possible response categories (e.g., correct and incorrect). One could say that both the person and the item have a position on the same latent trait and that position is characterized by a value along the real line. The probability that the person solves the item depends on the difference between the value of the person and the value of the item. Suppose that the person has latent value θ_v and that the item has latent value β_i, then the probability of an answer x_{vi} is

$$\Pr(X_{vi} = x_{vi} \mid \theta_v, \beta_i) = \frac{\exp[x_{vi}(\theta_v - \beta_i)]}{1 + \exp(\theta_v - \beta_i)}, \quad (1)$$

for $i = 1,...,k$ and $v = 1,...,n$ (i.e., a test with k items given to n persons). In Equation (1), x_{vi} is the realization of the random variable X_{vi} and takes the value 1 if the item is solved correctly and 0 otherwise. Note that if $X_{vi} = 0$, the numerator in Equation (1) reduces to 1.

An important assumption of this and most IRT models is that given the latent trait scores, every covariation between manifest indicators disappears. This assumption is called local stochastic independence (LSI). The adjective "local" refers to the fact that the independence only holds given the value of the latent traits. LSI is an important assumption in IRT models because it actually states that all the covariation between the item responses can be explained by means of a smaller set of latent variables. LSI means that once the latent trait score of the person is known, one also knows the probability that the person answers an item correctly, without the responses on the other items giving any additional information for this probability. This can be formulated in the following formal way:

$$\Pr(X_{vi} = x_{vi} \mid \theta_v, \beta_i, x_{v1}, \ldots, x_{v,i-1}, x_{v,i+1}, \ldots, x_{vk}) = \Pr(X_{vi} = x_{vi} \mid \theta_v, \beta_i). \quad (2)$$

If the assumption of LSI is violated in a data set, then there are remaining dependencies between the items after controlling for the latent traits. These dependencies are called local item dependencies (LIDs). In general, there are three possible causes for those LIDs. First, LIDs may be due to ignored multidimensionality. By multidimensionality, it is meant that more than one latent trait for the person is needed to explain the covariation among the items. Suppose that for solving the questions in a test two abilities are required: a numerical and a verbal ability. Moreover, suppose that a varying combination of the two traits is needed over questions, that is, for some questions the

numerical ability is the most important one for solving the question, while for other questions it is the verbal ability. If one neglects the fact that two abilities are necessary in the solution process and one tries to fit a unidimensional IRT-model to data of this kind, the resulting latent trait is a combination of the verbal and numerical abilities that is optimal in the entire set of questions. But, this one dimension cannot account for the more complex pattern of covariation between the items. Hence, there remains some dependency between the more verbal and between the more numerical items even after conditioning on the latent trait. In this paper we will not consider LIDs that are due to ignored multidimensionality.

Second, Differential Item Functioning (DIF) can also result in LIDs. DIF means that there is a relation between the responses and the membership of a group and the latent trait cannot explain this relation (Mellenbergh, 1985). After controlling for the estimated latent trait, items remain related to each other because of group membership. Also this type of LIDs will not be studied in this paper.

A final reason for the appearance of LIDs is that there are so-called item dependencies. This means that there are relations between some items that do not result from their mutual dependency on the latent trait(s). Three short examples will be given of how LIDs can show up in different contexts.

First, suppose that in a given test, groups of items can be discerned based on the content they refer to. For example, in a reading comprehension test, items typically refer to a given text. The items related to a common text may show more dependence than is expected from some standard IRT model that is used to analyze the data. The dependency may be caused by the way a particular text is read and understood, which affects all items referring to the text, beyond the underlying general reading comprehension ability.

As a second example suppose that solving an item correctly involves two components and that each of the components can be operationalized in a different item, called a subtask. Since the subtasks are part of the same total task they may be expected to interact. Hoskens and De Boeck (1995, 1997) have shown that when the subtasks are presented as separated items, the dependency is smaller than when the subtask responses are observed within the context of the total task.

Finally, consider a test in which the examinee gradually learns through a positive feedback mechanism. This implies that a correct answer on an item gives the examinee some knowledge about how to solve the subsequent items. Also in that case, the assumption of LSI does not hold.

These three examples are not meant to form an exhaustive list of situations in which item dependencies can show up. It should also be mentioned that mathematically spoken, this type of LIDs can be reformulated as a special case of ignored multidimensionality. However, looked upon from a more substantial

point of view, dependencies between items is something different than ignored multidimensionality, because no different additional latent traits for the person are brought into the model to explain the covariation among the items.

In the following sections, the issue of LIDs is further explored. First, psychometric models are presented in which the phenomenon of LID is translated in a formal model. Second, it is explained why these models can be estimated with standard statistical software like SPSS and SAS. Third, an application of the models is given for the analysis of response preferences. In the final section, some unresolved problems are discussed together with some suggestions for further research.

Psychometric Models for Local Item Dependencies

It is possible to formulate IRT models that can account for violations of LSI. Among others (e.g., Jannerone, 1986; Kelderman, 1984; Verhelst & Glas, 1993), Hoskens and De Boeck (1997) described models for LIDs, and these models will be explored first. Next, a more general model is discussed. In the following, the concepts "LIDs" and "(item) interactions" are used interchangeable. If items exhibit a form of local dependence, one could also say that they are interacting. This can be compared with an ANOVA framework: independent variables that do not interact, are not dependent on each other with respect to the effect they have.

Models for LIDs from Hoskens and De Boeck (1997)

In their conceptual analysis of different types of LIDs, Hoskens and De Boeck (1997) distinguished two classification dimensions: the type of interaction and the modus of interaction.

With respect to the type of interaction, there is a distinction between combination and order interaction. *Combination interaction* refers to the situation in which solving two items together involves something else (i.e., it is more easy or more difficult) than what can be expected from just solving the separate items. An example of this type of dependency occurs when two items in a test are linked to the same text. If one has the knowledge to solve one item, it increases the probability of solving also the other, because the two items partially overlap as they refer to the same text. On the other hand, if there is *order interaction*, then there is an order of the items to be discerned, with items that precede others having an effect on those others. The order can refer to the order of presentation, a logical order or a developmental order. Order interaction may show up when one is learning during the test taking and this depends on the order of presentation. In this case is solving an item informative about how

to handle the type of items of the test and that increases the probability of solving the following items of the same type.

With respect to the modus of the interaction, Hoskens and De Boeck (1997) distinguish between constant and dimension dependent interaction. If the interaction between the items is *constant*, this means that the strength and the direction (positive or negative) of the interaction is the same for all persons, no matter what their position on the latent trait is. In the *dimension dependent* interaction case, the strength and the direction of the interaction depends on the position of the person on the latent trait. Dimension dependent interaction may for example imply that for persons with a high latent trait value, the interaction is very strong and positive (e.g., a correct answer on an item increases the probability of a correct answer on a subsequent one), but that the strength of this interaction decreases together with the latent trait value. At a certain point on the latent continuum, the reverse effect appears: persons with a latent trait value lower than this point will show a negative interaction (e.g., a correct answer on an item decreases the probability of a correct answer on a subsequent one) and this tendency increases if the latent trait value further decreases.

Crossing the two classification dimensions results in four different models for item dependencies (or interactions between items). In Table 1, one can find the labels that are henceforth used to denote these four models that originate from the crossing of the two classification dimensions.

Table 1. *Four Models for LIDs*

	Type of the interaction	
Modus of the interaction	Constant combination Dimension dependent combination	Constant order Dimension dependent order

Before giving the mathematical formulation of the four interaction models, first the probability formula for the case of independence will be given to provide a baseline model for further comparison. If two items I_i and I_j are of the Rasch type and they are independent, the probability of a joint response (x_{vi}, x_{vj}) of person S_v equals:

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i, \beta_j) =$$

$$= \Pr(X_{vi} = x_{vi} \mid \theta_v, \beta_i) \Pr(X_{vj} = x_{vj} \mid \theta_v, \beta_j).$$

$$= \left(\frac{\exp[x_{vi}(\theta_v - \beta_i)]}{1 + \exp(\theta_v - \beta_i)} \right) \left(\frac{\exp[x_{vj}(\theta_v - \beta_j)]}{1 + \exp(\theta_v - \beta_j)} \right)$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i) + x_{vj}(\theta_v - \beta_j)]}{1 + \exp(\theta_v - \beta_i) + \exp(\theta_v - \beta_j) + \exp(2\theta_v - \beta_i - \beta_j)} \qquad (3)$$

From Equation (3) one sees that the joint probability (given the latent trait values) of the responses on two Rasch items equals the product of the marginal probabilities on the separate items (again given the latent trait values) and hence that there is LSI. This local independence relation holds for every vector of joint responses in the Rasch model.

The first model for LIDs that will be discussed is the *constant combination interaction model*. According to this model, the probability of a joint response (x_{vi}, x_{vj}) given the latent trait values is

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i, \beta_j, \beta_{ij}) =$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i) + x_{vj}(\theta_v - \beta_j) + x_{vi}x_{vj}(-\beta_{ij})]}{1 + \exp(\theta_v - \beta_i) + \exp(\theta_v - \beta_j) + \exp(2\theta_v - \beta_i - \beta_j - \beta_{ij})}$$

$$\neq \Pr(X_{vi} = x_{vi} \mid \theta_v, \beta_i) \Pr(X_{vj} = x_{vj} \mid \theta_v, \beta_j). \qquad (4)$$

As can be seen from Equation (4) in comparison with Equation (3), an interaction parameter β_{ij} is inserted in the probability formula. The parameter β_{ij} quantifies the interaction or local dependence between the items I_i and I_j. If β_{ij} is negative, this means that, for every θ_v the probability of a joint response $(X_{vi} = 1, X_{vj} = 1)$ will increase in comparison to the probability of the same joint response under the Rasch model. In this case, one could say there is positive interaction or dependence, because the covariation between the two items is higher than under the independence model. The reverse happens if β_{ij} is positive: for every θ_v the probability of a joint response $(X_{vi} = 1, X_{vj} = 1)$ will decrease in comparison to the probability of the same joint response under the Rasch model. In panel (a) of Figure 1 the probability for each possible joint response on two items I_1 and I_2 is drawn as a function of the latent trait value for the Rasch model and for the constant combination interaction model. The

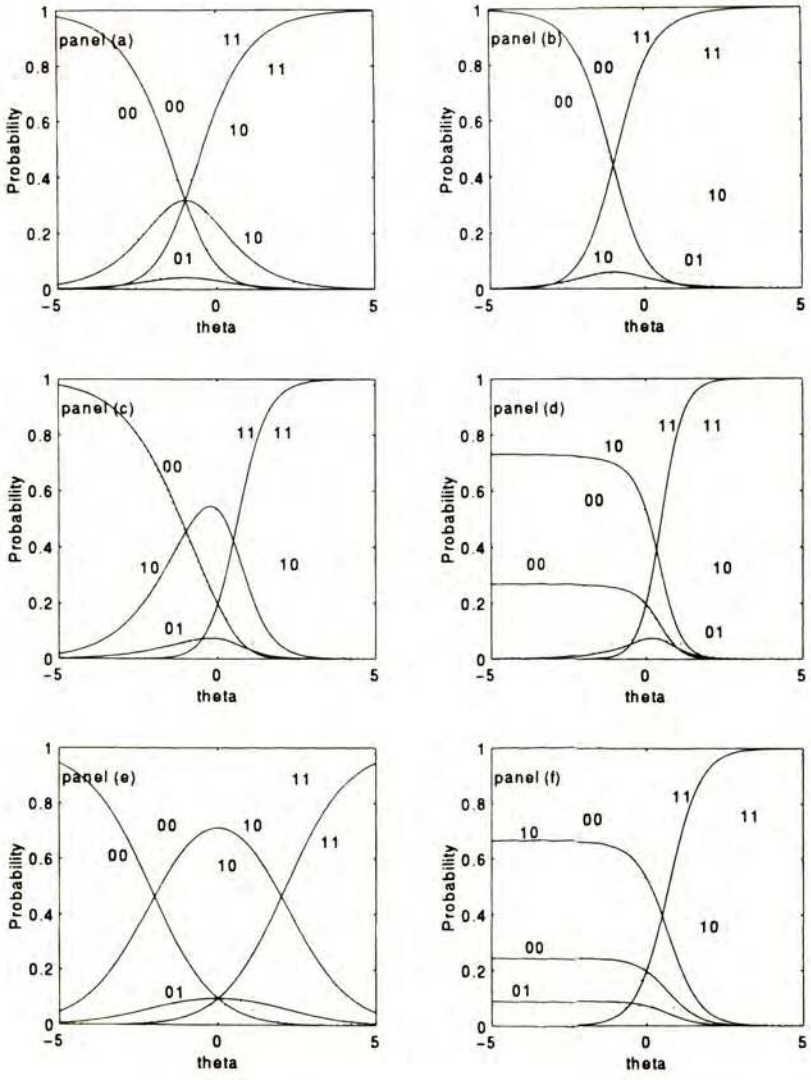

Figure 1. The probability of joint responses for (a) the constant combination interaction model (–), (b) the constant order interaction model (–), (c) the dimension dependent combination interaction model (–), (d) the dimension dependent order interaction model (–), (e) the constant alternation model (–) and (f) the dimension dependent alternation model (–) and in all panels (a)-(f) the Rasch model (...).

item parameters were chosen as follows: $\beta_1 = -1$, $\beta_2 = 1$ (this is true for both models) and $\beta_{12} = -2$ (for the interaction model). The interaction is positive (because β_{12} is negative) and it is seen that the probability of two correct responses on the two items is increased in comparison with the Rasch model. From Equation (4), it is also clear that the property of LSI does not hold anymore for the constant combination interaction model, except when β_{12} equals 0 as the models then become equal to each other.

A second model for LIDs is the constant order interaction model. According to this model, the probability of a joint response (x_{vi}, x_{vj}) given the latent trait values equals

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i', \beta_j', \beta_{ij}') =$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i') + x_{vj}(\theta_v - \beta_j') + x_{vi}(-1)^{1-x_{vj}}(-\beta_{ij}')]}{1 + \exp(\theta_v - \beta_i' + \beta_{ij}') + \exp(\theta_v - \beta_j') + \exp(2\theta_v - \beta_i' - \beta_j' - \beta_{ij}')}. \quad (5)$$

The primes for the item parameters in Equation (5) are used to denote that the same values for the parameters in this model and the constant combination interaction model do not lead to the same probabilities. Although, in the two item case, the constant order interaction model is only a reparametrization of the constant combination interaction model. Hoskens and De Boeck (1997) show that by equating Formulas (4) and (5), the following holds: $\beta'_i = \beta_i + \beta_{ij}/2$, $\beta'_j = \beta_j$ and $\beta'_{ij} = \beta_{ij}/2$.

Equation (5) closely resembles Equation (4) except for the last term in the numerator. This term equals 0 when $X_{vi} = 0$, but when $X_{vi} = 1$, the interaction parameter plays a role. However, β'_{ij} affects the probability of the joint responses $(X_{vi} = 1, X_{vj} = 1)$ and $(X_{vi} = 1, X_{vj} = 0)$ in an opposite direction. If β'_{ij} is negative, the probability of the response $(X_{vi} = 1, X_{vj} = 1)$ goes up and that of $(X_{vi} = 1, X_{vj} = 0)$ goes down. The reverse happens when β'_{ij} is positive. A graphical illustration of this constant order interaction model is shown in panel (b) of Figure 1. The parameters have exactly the same numerical values as in panel (a), but because of the differences between Equations (4) and (5), the form of the curves has changed. The negative interaction parameter implies that the probability of the joint correct response $(X_{v1} = 1, X_{v2} = 1)$ goes up for all latent trait values at the cost of the probability of the $(X_{v1} = 1, X_{v2} = 0)$ response, which becomes very small. (In this particular case the probability functions of the joint responses $(X_{v1} = 1, X_{v2} = 0)$ and $(X_{v1} = 0, X_{v2} = 1)$ are the same for constant order interaction model.)

The third interaction model is the *dimension dependent combination interaction model*. The probability of a joint response (x_{vi}, x_{vj}) given the latent trait values is

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i, \beta_j, \beta_{ij}) =$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i) + x_{vj}(\theta_v - \beta_j) + x_{vi}x_{vj}(\theta_v - \beta_{ij})]}{1 + \exp(\theta_v - \beta_i) + \exp(\theta_v - \beta_j) + \exp(3\theta_v - \beta_i - \beta_j - \beta_{ij})}. \quad (6)$$

In this model, it is not the value of β_{ij} anymore that matters solely, but the difference θ_v - β_{ij}. Hence, the direction and the magnitude of the item interaction varies depending both on the sign and on the absolute value of the difference θ_v - β_{ij}. For some persons the probability of (X_{vi} = 1,X_{vj} = 1) will increase and the probability of the other response patterns will decrease, while for other persons the reverse happens. Panel (c) of Figure 1 shows the probability curves for a dimension dependent combination interaction model with parameters β_1 = -1, β_2 = 1 and β_{12} = 0. It can be seen that the probability line for the joint response (X_{v1} = 1,X_{v2} = 1) under the dimension dependent combination interaction model crosses the one from the Rasch model. This means that for the model with these particular parameter values the high ability persons have a larger probability of two correct responses than under the Rasch model, while the low ability students have a lower probability of two correct responses than under the Rasch model.

The fourth and last model is the *dimension dependent order interaction model*:

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i', \beta_j', \beta_{ij}') =$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i') + x_{vj}(\theta_v - \beta_j') + x_{vi}(-1)^{1-x_{vj}}(\theta_v - \beta_{ij}')]}{1 + \exp(-\beta_i' + \beta_{ij}') + \exp(\theta_v - \beta_j') + \exp(3\theta_v - \beta_i' - \beta_j' - \beta_{ij}')}. \quad (7)$$

As for the constant order interaction model in Equation (5), the probability of the joint responses (X_{vi} = 1,X_{vj} = 1) and (X_{vi} = 1,X_{vj} = 0) are affected in an opposite way for the same β'_{ij} value in comparison with the Rasch model, but the amount and the direction of change of the probabilities now depends on the latent trait value of the person. If $\theta_v > \beta'_{ij}$, the probability of (X_{vi} = 1,X_{vj} = 1) will increase and the probability of (X_{vi} = 1,X_{vj} = 0) will decrease, with the amount of change dependent on the exact value of the difference, whereas the reverse happens when $\theta_v < \beta'_{ij}$. It should be noticed that in the dimension dependent case, combination and order interaction are not just reparametrizations of each other (even not for two items). Panel (d) of Figure 1 contains the probability curves of the four answer patterns for this model with the same parameter values as in the dimension dependent combination interaction case. The most salient feature from panel (d) is that the probability of (X_{v1} = 1,X_{v2} = 0) is

proportional to the probability of ($X_{v1} = 0, X_{v2} = 0$). The consequence is that under the dimension dependent order interaction model, giving a response equal to ($X_{v1} = 1, X_{v2} = 0$) is a sign of a low ability. For low ability persons, the response pattern ($X_{v1} = 1, X_{v2} = 0$) is much more probable than the response pattern ($X_{v1} = 0, X_{v2} = 0$) according to the dimension dependent order interaction model with these parameter values.

The concept of dimension dependent interaction is new in the literature. Although it can have a clear substantial interpretation for real data (examples can be found in Hoskens & De Boeck, 1995, 1997), it could be questioned whether dimension dependent interaction can be statistically differentiated from constant interaction, that is, whether the statistical tests have enough power to distinguish both types of interactions. This question can be formulated as the question under which conditions a true dimension dependent interaction model can be found back with data from a sample of finite size. To address this question, Tuerlinckx and De Boeck (in press) set up a simulation study. They showed that when the interaction parameter β_{ij} (from the dimension dependent interaction model) is not too extreme, the dimension dependent interaction model can be distinguished very well from the constant interaction model. However, when β_{ij} is very small or very large, most persons have a latent trait value that is respectively larger or smaller than the interaction parameter, meaning that the persons do not differ anymore with respect to the direction of the interaction, but only with respect to the relative magnitude of the interaction. Suppose for instance that in a dimension dependent combination interaction model β_{12} is -5 and θ_v follows a standard normal distribution. Then for most θ_v values, the difference between θ_v and β_{12} is positive, making the probability of ($X_{v1} = 1, X_{v2} = 1$) more probable than under the Rasch model for practically everyone. Hence, the data generated by this model will resemble the data generated by a constant combination interaction model. This is not the case when β_{12} has a moderate value. The same line of reasoning holds for a large positive value of β_{12}.

A Generalization of the Models of Hoskens and De Boeck (1997)

A more general model for LIDs, of which the previous ones are special cases is the following

$$\Pr(X_{vi} = x_{vi}, X_{vj} = x_{vj} \mid \theta_v, \beta_i, \beta_j, \beta_{ij}) =$$

$$= \frac{\exp[x_{vi}(\theta_v - \beta_i) + x_{vj}(\theta_v - \beta_j) + f(x_{vi}, x_{vj})(a\theta_v - \beta_{ij})]}{D}. \qquad (8)$$

with

$$D = \exp[f(0,0)(a\theta - \beta_{ij})] + \exp[\theta_v - \beta_i + f(1,0)(a\theta_v - \beta_{ij})]$$

$$+ \exp[\theta_v - \beta_j + f(0,1)(a\theta_v - \beta_{ij})] + \exp[2\theta_v - \beta_i - \beta_j + f(1,1)(a\theta_v - \beta_{ij})].$$

Two new things appear in Equation (8). First, the constant a is included, which can vary between 0 or 1. The two boundary values 0 and 1 correspond with constant and dimension dependent interaction, respectively. The smaller the a, the less dimension dependent the interaction is. Of course one could turn this constant a into a parameter α which has to be estimated from the data. This is not done in the present paper because it would make the estimation process much more complicated since α has to lie between 0 and 1. Second, Equation (8) contains the function f with arguments x_{vi} and x_{vj}, being the responses given on the items I_i and I_j by person S_v. In the case of combination interaction, this function is the conjunction or product of x_{vi} and x_{vj}, in the case of order interaction, it is defined as $x_{vi}(-1)^{1-x_{vj}}$. (Although the status of the β parameters changes with the exact definition of the function f, we did not indicate this in the formula.)

Of course, other functions than the ones just presented may be chosen. For instance, in the application section, a model will be estimated for which the function f is defined as

$$f(x_{vi}, x_{vj}) = XOR(1 - x_{vi}, x_{vj}) = 1 - x_{vi} + x_{vj} - 2(1 - x_{vi})x_{vj}. \qquad (9)$$

The function *XOR* denotes the exclusive 'or' and its arguments are binary valued variables. The value of the *XOR* function equals 1 only for the response patterns ($X_{vi} = 1, X_{vj} = 1$) and ($X_{vi} = 0, X_{vj} = 0$), and 0 for the other two response patterns. The model with this particular function for the interaction term is called the *alternation model*, because if $a = 0$ and β_{ij} is positive, the probabilities of the joint responses ($X_{vi} = 1, X_{vj} = 0$) and ($X_{vi} = 0, X_{vj} = 1$) increase in comparison with the Rasch model and the probabilities of the joint responses ($X_{vi} = 0, X_{vj} = 0$) and ($X_{vi} = 1, X_{vj} = 1$) go down. If a person would answer exactly according to a constant alternation model with a large β_{ij}, his or her response patterns would consist of alternating responses. It can be shown that a constant alternation model is just a reparametrization of the previously presented constant interaction models. Furthermore, if $a = 1$, one has a dimension dependent alternation model for which some persons (if $\theta_v - \beta_{ij} < 0$) alternate between the responses and other persons (if $\theta_v - \beta_{ij} > 0$) answer more consistently with the same response. In both cases, the degree of showing one or the other response tendency depends on the difference between the latent trait value of the person and that of the item. For the case of dimension dependent alternation, the model

is not equivalent anymore to the previous presented dimension dependent models.

In Figure 1 panel (e), one can see the probability curves of the constant alternation model for the joint responses in case there are two items and with item parameters equaling: $\beta_1 = -1$, $\beta_2 = 1$ and $\beta_{12} = 1$. It is clear from the graph in panel (e) that the probability of the alternating responses go up in comparison with the Rasch model. Panel (f) of Figure 1 shows the dimension dependent alternation model. The probability curves of the joint responses ($X_{v1} = 1, X_{v2} = 0$) and ($X_{v1} = 0, X_{v2} = 1$) are proportional to the one of ($X_{v1} = 0, X_{v2} = 0$), meaning that persons with a low latent trait value will tend to give relative more alternation responses (of both types) in comparison with the Rasch model. Persons with a high latent trait value will tend to give more ($X_{v1} = 1, X_{v2} = 1$) responses than under the Rasch model.

Estimating and Testing the Models for Item Dependencies

Although the presented models look complex, they can be estimated and tested rather easily in standard statistical software packages like for example, SPSS or SAS. In this section it will be explained why standard statistical software can be used rather than showing how it can be done. Details about the latter subject can be found in Tuerlinckx (1996). In the discussion section, a number of disadvantages of this classical approach of estimating and testing the models are given together with some solutions.

The most natural way to estimate the parameters of the models would be to maximize the likelihood which is a function of all the parameters in the model (item parameters as well as person parameters); this likelihood is called the joint likelihood. When one increases the number of examinees that take the test, one would expect to have an estimate of the item parameters that comes closer to its true value. However, this property of consistency does not hold when one maximizes the joint likelihood: increasing the number of persons also increases the number of parameters as every person brings one parameter along, and this destroys the property of consistency. Consistency is an important property of estimators and therefore, this way of handling the estimation problem has been abandoned.

A nice feature of the models above is that the sum score of a person (i.e., the total number of correct responses) is a so-called minimal sufficient statistic for the ability parameter. This means that all the information about the ability parameter is captured in this sum score and no other function of the data (even not the data themselves) will give more information about the ability parameter than the sum score. A consequence is that conditional on the sum score of the person, the probability of a correct response is only a function of the item

parameters and not of the person parameters anymore, because everything that can be known about the subject parameters is known through the conditioning on their (minimal) sufficient statistics. Since it is our purpose in the first place to gain knowledge about the dependencies between items, we are not interested in the ability parameters, and hence, dropping the person parameters by conditioning on the sum score is not a problem. If one wants to make inferences about the person parameters, they can be estimated after the item parameters are estimated.

Because the probabilities of the response patterns given the sum score are only function of the item parameters, one can estimate those item parameters and test the model with a loglinear analysis. An explanation of the theory of loglinear models can be found in Agresti (1990), but a grasp of it can be given by drawing an analogue with the well-known ANCOVA model. In an ANCOVA framework, the expected value of the continuous dependent variable is decomposed into main effect parameters (one for each independent variable) and interaction parameters, while also a continuous covariate is inserted to adjust the expected value of the dependent variable for non-manipulated but measured continuous effects that are assumed to have an influence on the dependent variable.

Loglinear models can be compared with an ANCOVA model, given some necessary adaptations. In three points these necessary adaptations will be given and applied directly to the special case of item dependencies in IRT. (1) It is the logarithm of the expected frequency of a response pattern given a sum score that serves as the dependent variable and will be decomposed in a similar way as for the ANCOVA model. The expected frequency is obtained by multiplying the probability of a response pattern (given the sum score) by the number of persons n. (2) The experimental conditions (or main effects) are the items. In the case of binary items, each condition has two levels, and hence, one free parameter is brought into the model by a condition or item. This free parameter is the item parameter. The interaction parameters stem from the interaction between experimental conditions. (3) The sum score is not used as such in the model but it is transformed into binary valued variables and the number of those variables equals the number of possible sum scores. These binary variables function as covariates.

An example of such a decomposition for three items is shown in Table 2 and it is the loglinear equivalent of a constant combination interaction model for items I_1 and I_2 and a Rasch model for item I_3. The γ parameters correspond to the β parameters of the original IRT model, but they differ from the original β parameters by a linear transformation. It can be seen that the sum score (which varies from 0 up to 3) is transformed into 4 binary variables (s_0, s_1, s_2 and s_3) which are present in the decomposition if the sum score corresponds with their index and are left out otherwise. In Table 2 there is also a general constant that

is added to all decompositions. This constant term denotes the grand mean (of the logarithm of expected frequencies). A mathematical derivation of the loglinear formulation of the Rasch model and of the interaction models for item dependencies can be found in Kelderman (1984) and in Tuerlinckx (1996), respectively.

Table 2. *An Example of a Loglinear IRT Model*

Response pattern	Sum score	Decomposition
000	0	$\mu + s_0$
100	1	$\mu + \gamma_1 + s_1$
010	1	$\mu + \gamma_2 + s_1$
001	1	$\mu + \gamma_3 + s_1$
110	2	$\mu + \gamma_1 + \gamma_2 + \gamma_{12} + s_2$
101	2	$\mu + \gamma_1 + \gamma_3 + s_2$
011	2	$\mu + \gamma_2 + \gamma_3 + s_2$
111	3	$\mu + \gamma_1 + \gamma_2 + \gamma_3 + \gamma_{12} + s_3$

Three additional things have to be noted with respect to Table 2. First, an interaction term is included between items I_1 and I_2, but the interaction term only appears when X_{v1} and X_{v2} are both one. Actually, there are four different interaction terms, one for each possible combination of X_{v1} and X_{v2}. But like in an ANCOVA model, only one of those interaction terms is a free parameter; hence, the other can be put to zero and they disappear from the decomposition formulas (dummy coding). Second, the set of parameters as shown in Table 2 cannot be estimated because the model is not identified. Some additional restrictions have to be put on the parameters: $\gamma_3 = 0$ (γ_3 is the parameter for item I_3) and $s_3 = 0$ (s_3 is the parameter for sum score 3), for example. We will not go into details about the exact nature of these restrictions. The interested reader is referred to Tuerlinckx (1996). Finally, since the model presented in Table 2 is the loglinear equivalent of a constant combination interaction model for items I_1 and I_2 and a Rasch model for item I_3, no interaction terms are included for interactions between I_1 and I_3, between I_2 and I_3, and between all three items.

After estimating the parameters of a model, the next step in the process of applying a statistical model, is checking how well it fits the data. If, as outlined above, the models are set up within a loglinear framework, measures of badness-of-fit are the usual Pearson chi-square and likelihood ratio chi-square test statistics (Agresti, 1990). Under the null hypothesis ("the tested model is true"), the distribution of these statistics is asymptotically chi-square with degrees of freedom equal to the number of free cells minus the number of free parameters in the model.

An Application: Examining the Structure of the Polish Questionnaire

As an illustration of the above modeling approach, we will apply the general model to a situation that pertains to the issue of response preferences. In his overview article, Tune (1964) showed that in different research areas in psychology a sequence of responses given by persons generally showed the same characteristics, namely that sequences are nonrandom but contain interdependencies. This is for example the case when persons have to judge whether they perceive a sequence of stimuli when those stimuli are in fact not present, or below the absolute threshold, or when persons are asked to generate sequences of random numbers, the output does not coincide with what is generally seen as random (Wagenaar, 1972). The same is also found when people are performing in probability learning tasks (Tune, 1964) in which people have to predict the next outcome of a stochastic process.

The main feature of these non-random sequences is that there is a tendency to alternate too much between the possible outcomes in comparison with what can be expected if the sequence were really random. In the literature about generating random numbers this phenomenon is called the *negative recency effect*, in the gambling literature it is called the *gambler's fallacy*. Gamblers believe that there exists some self-correcting mechanism in the generation process of random numbers so that if the same event has occurred for some time, the probability of the other non-occurred events will increase. In other contexts this phenomenon is called the contrast effect. Budescu (1987) reports individual differences in that tendency to alternate, that is, some persons alternate, others just a little bit while some do the reverse and respond quite consistently. For the latter ones, there is a positive dependency or a positive correlation between subsequent responses.

In this section, a new research method is used to investigate the tendency to alternate. Persons had to fill out a so-called Polish questionnaire. In the instructions they were told that they had to answer yes or no to 10 questions from a Polish intelligence test to see whether common language structures between Polish and Dutch had an influence on the accurateness of their responses. However, the questions were not intelligence questions, but they were non-sentences consisting of only parts of different Polish sentences pasted together. We believe that this situation has some resemblance with situations where one has to judge whether an auditory signal is present when there is none, or predicting whether a light will go on or not (at least when there is no feedback) and therefore, that the paradigm can be used to study response preferences.

IRT models for item dependencies were used to analyze the data, given that we expected to find differences in the generated sequences and dependencies among the responses. From a theoretical point of view, the dimension dependent alternation model with interactions between subsequent items seems a

plausible and quite simple model. First, it explains the alternation and interindividual differences of two kinds (in alternation -as found by Budescu, 1987- and in proportion yes responses). Suppose we are in the situation that is shown panel (f) of Figure 1. Persons with a large positive θ_v -value will tend to give consistent ($X_{v1} = 1, X_{v2} = 1$) response patterns that also result in a high proportion of yes responses. On the other hand, persons with a large negative θ_v -value will tend to give mostly inconsistent (or alternating) ($X_{v1} = 1, X_{v2} = 0$) response patterns that result in a moderate proportion of yes responses. Second, if the model fits, this means that if a certain response is given on item I_i, it implies a lower chance of giving the same response on item I_j for persons with $\theta_v - \beta_{ij} < 0$ (see also Figure 1, panel (f)). It is also assumed this is the case for the majority of the persons (consequently β_{ij} has to be large enough) and that only a few show the reverse effect, since it is expected that most persons show a negative recency effect. The latent trait can be seen as a kind of alternation-consistency response trait where persons with a low latent trait value alternate a lot, while persons with a high latent trait value have consistent response patterns.

Participants

The participants were 465 first year law students from a Flemish university. They had to fill out the Polish questionnaire. For 26 participants, one or more responses were missing and the responses of these persons were not included in the analysis. Sixteen participants indicated they spoke Polish, however none of them remarked the fact that non-sentences were presented. Therefore, these persons were nevertheless included in the analysis.

Method

The IRT models for item dependencies were fitted to the data by means of a loglinear model analysis with SPSS. However, in using loglinear analysis one is restricted with respect to the numbers of items. Analyzing all 10 items together using loglinear models would make the estimation and testing procedures unreliable due to a large amount of low and zero frequencies in the table. Hence, the analyses were done separately for the first five and last five items. We will come back to this problem in the discussion section.

Results

Descriptive summary. Before showing the results of the IRT analysis, some more classical statistics are computed for the ten items. The proportion of yes

responses per person has a distribution with mean 0.525 ($SD = 0.150$). Table 3 shows the frequency distribution of proportion yes responses. Some persons responded 0 or 1 to every item; probably, they used a strategy to obtain a good score at the so-called intelligence test. The probability Pr (A) is the probability of alternation and it is computed as the actual number of alternating pairs of adjacent responses (01 or 10) divided by the maximum number of pairs of adjacent responses in the sequence of 10 responses. The obtained value is 0.611 ($SD = 0.217$), which is a rather high value. Such a high value is consistent with the findings of other authors in the literature. For example, Falk and Konold (1997) report the probability of alternation from nine different studies in which people had to generate a random sequence; the mean value of Pr (A) over the nine studies is 0.593.

Table 3. *Summary Statistics*

p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
N	12	2	4	9	64	161	132	43	3	1	8

Psychometric analysis. The fitted models for the first five items are presented in Table 4. An identification number for the different models is shown in the first column. The second column contains labels for the fitted models; these labels will be clarified further on in the paper. The third column has in its rows the two proposed test statistics, the likelihood ratio chi-square (L^2) and the Pearson chi-square statistic (X^2), and the next column indicates the degrees of freedom for both test statistics. Their p-value is shown in the fifth column and in the last column, one finds the *AIC* (Akaike's information criterion; Akaike, 1977). The *AIC* is a badness-of-fit value that can be used to compare the fit of non-nested model. Two models are called non-nested models when none of them can be derived from the other by restricting one or more parameters to zero. For instance, the dimension dependent and the constant interaction models are non-nested models. The *AIC* adds a penalizing function for the number of parameters to the likelihood ratio chi-square statistic.

The first two rows in Table 4 refer to the results of fitting two standard loglinear models. Model 1 is an independence model for which it is assumed that the five items are independent. Model 2 assumes first order interactions of the alternation type between subsequent items, but without individual differences. All fit indices indicate that these models do not fit. Some improvement in fit was achieved by including the sum score as a variable in the independence model, which boils down to fitting a Rasch model (Model 3). The fit of the latter model was not sufficient, however. The better fit of Model 3 indicates that an

Table 4. *The Fitted Models for the First Five Items*

Nr.	Model	Test statistic	*df*	*p*-value	*AIC*
1	Independence	$L^2 = 258.950$	26	0.000	270.950
		$X^2 = 270.810$	26	0.000	
2	Subsequent interactions	$L^2 = 123.447$	22	0.000	143.447
		$X^2 = 157.465$	22	0.000	
3	Rasch model	$L^2 = 109.730$	22	0.000	129.730
		$X^2 = 116.352$	22	0.000	
4	Constant alternation	$L^2 = 31.105$	18	0.028	59.105
	(free)	$X^2 = 29.397$	18	0.044	
5	Constant alternation	$L^2 = 34.110$	21	0.035	56.110
	(restr.)	$X^2 = 33.501$	21	0.041	
6	Dimension dependent	$L^2 = 102.946$	18	0.000	126.946
	alternation (free)	$X^2 = 142.874$	18	0.000	
7	Constant alternation	$L^2 = 21.172$	18	0.271	49.172
	(Model 5 + lag 2 - free)	$X^2 = 22.132$	18	0.226	
8	Constant alternation	$L^2 = 22.254$	20	0.327	46.254
	(Model 5 + $I_2 \times I_4$)	$X^2 = 23.522$	20	0.264	

interindividual difference variable (here the sum score) was needed to account for the variance in the data, hence, also the following models took into account individual differences with respect to the proportion of yes responses.

Model 4 is a constant alternation model with interactions between the subsequent items. The term "free" between parentheses means that the magnitude of the interactions parameter were allowed to vary over item pairs. Including interactions in the model improved the fit quite drastically, but it was still not very good. The constant alternation model (Model 5) with equal interactions between subsequent items did not have a worse fit; the AIC was even smaller. However, the actual absolute fit of the model to the data was not satisfying enough to complete the analysis. But also the dimension dependent alternation model (Model 6) did not have a good fit to the data, and in fact the fit was even worse than for the constant alternation models.

Because all the previous simple models failed to fit the data, the models were expanded. Taking the constant alternation model with equal interactions (Model 5) as a starting point, three extra interactions were added. These interactions are between items that are at one item distance from each other: items I_1 and I_3, I_2 and I_4 and I_3 and I_5. This means that the lag of the interaction is increased with one unit from 1 to 2. In Model 7 these interactions were taken to be free and this gave a good fit to the data. The parameter estimates of this model showed that the common lag 1 interaction parameter was quite large

($\beta_{i,i+1} = 0.630$). Two of the three lag 2 interaction parameters did not differ very much from zero ($\beta_{13} = 0.073$ and $\beta_{35} = -0.067$), only the interaction parameter between the items I_2 and I_4 differed significantly from zero ($\beta_{24} = 0.390$). Leaving the two non-significant interactions out, the resulting model had a quite good fit (see Model 8). The remaining parameters of this model were almost equal to the ones of the previous model (Model 7). A model selection procedure based on the *AIC* pointed at this model too as the best fitting one.

Fitting the constant alternation model with the smallest *AIC* value to the last remaining five items that were not used in the previous analysis resulted in a bad fit to the data ($L^2 = 92.155$ and $X^2 = 89.068$, $df = 18$, $p = 0.000$). The interaction structure for the last five items of the test was very complex and we will not present the results of these analyses. Interactions between three and four items had to be included and those higher order interactions made the model difficult to interpret. Also, the interactions did not show anymore a clear preference for alternation between responses.

Conclusion

The hypothesis that the dependency relations for the first five items could be explained by a dimension dependent alternation model with subsequent interactions (Model 6) was rejected. Instead a constant alternation model with interactions between subsequent items and one lag 2 interaction fitted (Model 8) the first part of the data very well. This model indicates that there are no interindividual differences in the alternation tendency but only in the proportion of yes responses.

We suggest that the complex interaction pattern for the second half of the test is the result of the fact that a lot of people will adapt their response strategy to end up with approximately 50% of yes responses. Some indirect evidence for this post hoc explanation can be found by looking at the raw data. As a first indication, one sees from Table 3 that most of the participants ended up with an observed proportion of yes responses around 0.5 (for 81% of the participants the proportion of yes responses was between 0.4 and 0.6). Second, the correlation between the sum score on the first half of the questionnaire and the sum score on the second half was only 0.365. Moreover, the correlation between the number of alternations for the first five items and the number of alternations for the second five items was only 0.367. These two low correlations indicate that the persons change their way of answering in the second half compared to the first one and this could be a factor in the genesis of the complex interactions for the last five items.

Discussion

The analyses on the Polish Questionnaire indicated some deficiencies in the way the models are estimated and tested. These deficiencies can be called structural deficiencies as they result from reformulating the models for LIDs to loglinear models to make them estimable and testable.

A first drawback from the estimation procedure is that only a limited number of items can be used in the analysis. The computer programs that estimate loglinear models store all possible response patterns, together with their frequencies as a basis for the calculations. Many items require an enormous amount of computer memory, which is not available to most computers. For instance, with 20 binary items a total of $2^{20} = 1,048,576$ different response patterns must be stored in the computer memory.

A second drawback of loglinear models pertains to the global test statistics, L^2 and X^2, that are used to give an indication of the fit of the model to the data. The interpretation of both statistics relies on an asymptotic approximation to the χ^2 distribution. This approximation can only be assured when the number of observations is very large for each possible response pattern. In our application this is certainly not the case if we analyze all 10 items together when there are only 439 observations. The situation is clearly better when only five items are analyzed. Especially when the difference between the L^2 and X^2 statistics is large, the approximation can be questioned. For instance, in Table 4, there is a substantial difference between the two test statistics for the subsequent interaction model (Model 2), while only five items are used.

These two disadvantages of the loglinear estimation and testing procedure make it difficult to judge the true value of using the models for item dependencies, for instance, when investigating the structure of response preferences within a moderate set of items. However, that should not mean that those models cannot be used. Two solutions exist to overcome the problems.

The first solution is called marginal maximum likelihood (MML; Bock & Aitkin, 1981; Thissen, 1982; Verhelst, 1993) estimation. The core of this approach is not to take the probabilities of (joint) responses given the person and item parameter values as the base for one's inferences, but the probabilities of the (joint) responses only given the item parameter values. This implies that the person parameter values have to be integrated out of the probability functions and this requires that one specifies the density function of the person parameter values. Using MML estimation, one is not troubled anymore by the presence of the many latent person parameters. However, the consequence is that the model is expanded, because an assumption has to be made about the latent distribution of person parameters. Moreover, this approach is only a way out of the limitations of the traditional estimation procedure as it does not solve the difficulties of testing the proposed models. Also in the MML approach, one

has to rely on the chi-square based badness-of-fit measures.

A second solution is to adopt a Bayesian framework (Gelman, Carlin, Stern, & Rubin, 1995) instead of the classical statistical framework, which was followed in this paper. In the classical statistical framework (also called the frequentist approach), the parameters of the models are considered as fixed, meaning that they have a true, but unknown, value. The task of estimation is to find a good approximation for this unknown value. The frequentist's criteria for a good approximation can be found for example in Mood, Graybill and Boes (1974). Another possibility is to consider the parameters as random variables, which is the approach taken in Bayesian statistics. The first step in a Bayesian analysis is to set up a distribution for the parameters that reflects one's prior knowledge (or absence of knowledge) about the reasonable values for these parameters. This distribution is called the prior distribution. These prior distributions are adjusted by the likelihood (which contains the model and the data), and the result is a posterior distribution. The posterior distribution reflects the uncertainty about the parameters after the data collection. Using Bayesian statistics can help in estimating the model by exploring the posterior distribution. Testing is done by generating new data under the model (using a large sample from the posterior) and comparing the characteristics of these model-based data sets with the characteristics of the observed data. This is the technique of posterior predictive checks (Rubin, 1984). If the model-based replicated data do not resemble the observed data, the model is not correct for the given data set. This Bayesian approach has already been applied to IRT by Janssen, Tuerlinckx, Meulders and De Boeck (in press) in the analysis of attainment goals of primary education. However, applying Bayesian methods for the estimation and testing of item dependency models

remains a topics for future research.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Akaike, H. (1977). On entropy maximization principle. In P.R. Krishnaiah (Ed.), *Proceedings of the symposium on application of statistics* (pp. 27-47). Amsterdam: North-Holland.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM-algorithm. *Psychometrika, 46*, 443-459.

Budescu, D.V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 25-39.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Fischer, G.H., & Molenaar, I.W. (Eds.). (1995). *Rasch models: Foundations, recent*

developments, and applications. New York: Springer.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Hoskens, M., & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32*, 364-384.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependencies among test items. *Psychological Methods, 2*, 261-277.

Jannerone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (in press). An hierarchical IRT model for mastery classification. *Journal of Educational and Behavioral Statistics*.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Mellenbergh, G.J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek [Item bias: definition, detection and research]. *Nederlands Tijdschrift voor Psychologie, 40*, 425-435.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the theory of statistics*. London: McGraw-Hill.

Rasch, G. (1980). *Probabilistic models for intelligence and attainment tests*. Chicago: The University of Chicago Press.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12*, 1151-1172.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

Tuerlinckx, F. (1996). *Loglineaire formuleringen van modellen voor itemafhankelijkheden*. [Loglinear formulations of models for item dependencies]. Unpublished master's thesis, Department of Psychology, University of Leuven, Leuven, Belgium.

Tuerlinckx, F., & De Boeck, P. (in press). Distinguishing constant and dimension dependent interaction: A simulation study. *Applied Psychological Measurement*.

Tune, G.S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin, 61*, 286-302.

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

Verhelst, N.D. (1993). Itemresponstheorie. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 83-178). Cito: Arnhem.

Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395-415.

Wagenaar, W.A. (1972). Generation of random sequences by human subjects: a critical survey of literature. *Psychological Bulletin, 77*, 65-72.

Received March, 1998
Accepted September, 1998