Distinguishing Constant and Dimension-Dependent Interaction: A Simulation Study

Francis Tuerlinckx and Paul De Boeck
University of Leuven

A simulation study was conducted to determine how well two models for local item dependency (LID), called interaction models, could be distinguished. The models examined were the constant order interaction model (COIM) and the dimension-dependent interaction model (DDIM). Data were simulated according to the latter model. Three factors were manipulated: sample size, the weight of the difference between the latent trait value of the examinee and the interaction parameter, and the value of the interaction parameter. Results indicated that (1) if the interaction parameter is not too extreme, the COIM will be rejected in favor of the true model (the Rasch model fit poorly for all levels of the interaction parameter); (2) a larger weight of the difference between the latent trait value and the interaction parameter facilitated the rejection of the COIM, although finding the true weight required a large sample size; and (3) the value for the interaction parameter with an optimal discrimination between the COIM and DDIM was not 0, as expected. *Index terms: constant interaction, dimension-dependent interaction, item response theory, local item dependencies, local stochastic independence, order interaction.*

A central assumption in most psychometric models is local stochastic independence (LI). LI means that, given all relevant latent trait values ($\boldsymbol{\theta}$) of a person, the probability of any item response does not depend on the other item responses. This assumption is violated if dependencies beyond the latent traits exist between items. These dependencies are called local item dependencies (LIDs) or interactions between items, because the joint probability of a response pattern, given $\boldsymbol{\theta}$, cannot be predicted from the probabilities of the responses on single items. LIDs due to ignored multidimensionality are not considered here.

One way to account for LIDs is by means of psychometric models for testlets (Wainer & Kiely, 1987). A testlet is a set of items that belong together and are analyzed as such, considering the response patterns on the items that are modeled as response categories of a single item. Within a testlet, violations of LI are allowed, but LI is required between different testlets. Models for testlets are described by Wilson & Adams (1995), Hoskens & De Boeck (1997), Jannerone (1986), and Kelderman (1984). The approach of Hoskens & De Boeck is used here.

Types of LID

Hoskens & De Boeck (1997) distinguished conceptually between order and combination interaction. Models for order interaction focus on learning and carry-over effects: Due to a specific order, previous items have an influence on subsequent items. Combination interaction is related to situations in which solving a total task involves more than solving the separate subtasks. Hoskens & De Boeck (1997) also distinguished between constant and dimension-dependent interaction. In

Applied Psychological Measurement, Vol. 23 No. 4, December 1999, 299–307
©1999 Sage Publications, Inc.

a constant-order interaction model (COIM), the magnitude and the direction of the interaction are equal for all persons. In contrast with this model, Hoskens & De Boeck formulated a dimension-dependent interaction model (DDIM). In this model, the interaction depends on the position of the person on the latent trait that the items are intended to measure. A similar distinction has been made by Hoskens & De Boeck (1995) in the context of componential item response theory models.

The mathematical formulation of both interaction models is a generalization of the Rasch (1960/80) model. In principle, the two-parameter logistic model (2PLM) could be used as a starting point, but some results used here (primarily concerning estimation) do not hold in that case. However, the distinction between constant and dimension-dependent interaction also applies to the 2PLM, as explained by Hoskens & De Boeck (1997). Assume that a testlet consists of two dichotomously scored items, I_1 and I_2. First, the joint probability of the answer pattern (x_1, x_2) according to the Rasch model is

$$P(X_1 = x_1, X_2 = x_2 | \theta, \beta_1, \beta_2) = \frac{\exp[x_1(\theta - \beta_1) + x_2(\theta - \beta_2)]}{1 + \exp(\theta - \beta_1) + \exp(\theta - \beta_2) + \exp(2\theta - \beta_1 - \beta_2)} = P(X_1 = x_1 | \theta, \beta_1) P(X_2 = x_2 | \theta, \beta_2) \quad , \tag{1}$$

where θ represents the position of the person on the latent trait and β_i represents the position of item I_i. The COIM is a generalization of the model in Equation 1:

$$P(X_1 = x_1, X_2 = x_2 | \theta, \beta_1, \beta_2, \beta_{12}) = \frac{\exp\left[x_1(\theta - \beta_1) + x_2(\theta - \beta_2) + x_1(-1)^{1-x_2}(-\beta_{12})\right]}{1 + \exp(\theta - \beta_1 + \beta_{12}) + \exp(\theta - \beta_2) + \exp(2\theta - \beta_1 - \beta_2 - \beta_{12})} \; . \tag{2}$$

The parameter β_{12} is the interaction parameter, and it affects the probability of the response patterns $(1,1)$ and $(1,0)$ in opposite directions. When the probability of $(1,1)$ increases, the probability of $(1,0)$ decreases. If $\beta_{12} < 0$, this indicates the presence of positive feedback (or learning), i.e., when the first item is answered correctly, the probability that the second item will be answered correctly increases. If $\beta_{12} > 0$, the reverse is true, indicating negative feedback. This model accounts for a violation of LI, but only to the extent that this violation is equal over all persons. Note that if $\beta_{12} = 0$, Equation 2 simplifies to Equation 1, and hence LI holds.

The DDIM is

$$P(X_1 = x_1, X_2 = x_2 | \theta, \beta_1, \beta_2, \beta_{12}) = \frac{\exp\left[x_1(\theta - \beta_1) + x_2(\theta - \beta_2) + x_1(-1)^{1-x_2}(\theta - \beta_{12})\right]}{1 + \exp(-\beta_1 + \beta_{12}) + \exp(\theta - \beta_2) + \exp(3\theta - \beta_1 - \beta_2 - \beta_{12})} \; . \tag{3}$$

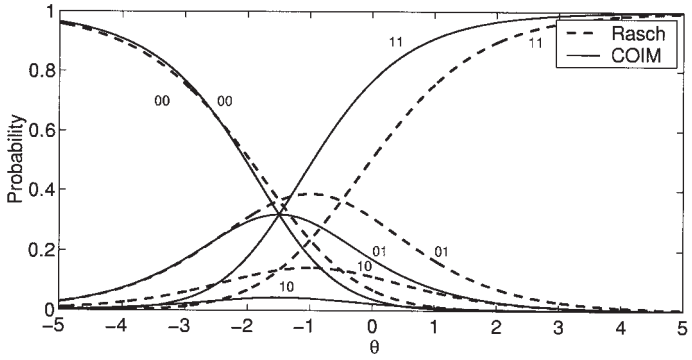
In this model, it is the value of $(\theta - \beta_{12})$ that is important. If this value is positive, the probability of a $(1,1)$ response will increase and the probability of a $(1,0)$ response will decrease; the amount of this change is dependent on the difference between θ and β_{12}. The reverse occurs if the value of $(\theta - \beta_{12})$ is negative.

Sample item response functions (IRFs) for the COIM and DDIM are plotted in Figure 1. They are compared with IRFs of response patterns for which the separate items correspond to the Rasch model. As Figure 1 shows, the Rasch model results in different IRFs than the two interaction models.

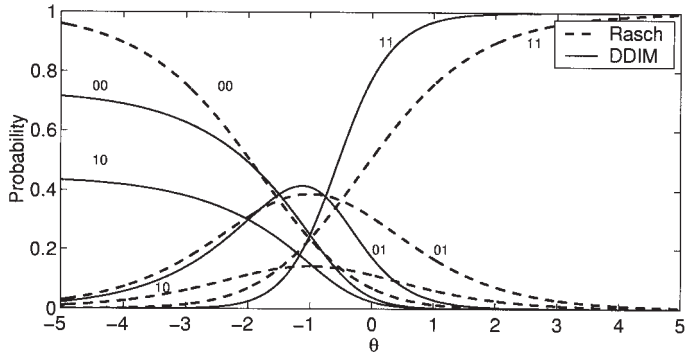
The models in Equations 2 and 3 are the simplest examples of interaction models—more complicated models can be constructed. For example, it is possible to weight the differences between person parameters and item parameters by adding a weight to the difference $(\theta - \beta_{12})$. These weights can be considered as known constants, and their appropriateness inferred from the fit of

Figure 1
IRFs for the Joint Probability of Answer Patterns for Two Items: $\beta_1 = -.5, \beta_2 = -1.5, \beta_{12} = -1$

a. Rasch Model and COIM



b. Rasch Model and DDIM



the model to the data. Or the weights can be considered as parameters, in which case the model approaches the more general 2PLM, because a discrimination parameter is added for one of its terms.

An Example of Dimension-Dependent Interaction

Suppose that individuals who take an ability test learn during the test by answering an item correctly. Also suppose that the learning ability depends on the position of the individual on the ability continuum. Then, persons with high ability will produce more correct responses, not only because their ability exceeds more item difficulties but also because their ability dominates over the interaction parameter. Evidence for this kind of dimension-dependent type of learning was found by Bell, Pattison, & Withers (1988). They observed that the odds ratio between two interacting items increased along with increasing ability. A constant interaction between two items would mean that the odds ratio remains relatively unchanged over ability levels.

However, this is only indirect evidence. More direct evidence supporting the model in Equation 3 can be found in an application by Hoskens & De Boeck (1997). They demonstrated that a DDIM fit the data well and that it fit better than a COIM.

Method

The present study was concerned with how well constant and dimension-dependent interaction could be distinguished in a simulated dataset. Data were generated according to the DDIM and analyzed with an independence model (the Rasch model), a COIM, and a DDIM. Several criteria of fit were used to determine the conditions under which these models were distinguishable from each other.

Datasets

The DDIM was used as the "true" or generating model for five items. The difficulty (β) parameters for the items were -1.25, $-.75$, 0.00, .75, and 1.25. Between each two adjacent items there was a dimension-dependent interaction, and this interaction was restricted to be equal; hence, $\beta_{12} = \beta_{23} = \beta_{34} = \beta_{45}$. θs were randomly drawn from a normal distribution with mean 0 and variance 1. See Park & Miller (1988) for the random number generator used.

Three factors were manipulated. The first factor was sample size (n = 100, 500, 5,000). The second was the value of the interaction parameter ($\beta_{ij} = -2, -.5, 0.0, .5, 2.0$). A small β_{ij} means that for a large majority of simulees (with $\theta > \beta_{ij}$), the probability of a $(1,1)$ response was increased and that of a $(1,0)$ response was decreased. With a large β_{ij} the reverse is true, i.e., for a large majority of simulees ($\theta < \beta_{ij}$) the probability of a $(1,1)$ response decreased, whereas their probability of a $(1,0)$ response increased.

Equation 3 can be rewritten for items I_i and I_j with a discrimination index or weight a_{ij} for the interaction term as

$$P(X_i = x_i, X_j = x_j \,|\theta, \beta_i, \beta_j, \beta_{ij}) = \frac{\exp\left[x_i(\theta - \beta_i) + x_j(\theta - \beta_j) + a_{ij}x_i(-1)^{1-x_j}(\theta - \beta_{ij})\right]}{1+\exp[(1-a_{ij})\theta - \beta_i + a_{ij}\beta_{ij}] + \exp(\theta - \beta_j) + \exp\left[(2+a_{ij})\theta - \beta_i - \beta_j - a_{ij}\beta_{ij}\right]}. \quad (4)$$

The third factor varied set $a_{ij} = 1$ in one-half of the conditions and $a_{ij} = 2$ in the other half ($a_{12} = a_{23} = a_{34} = a_{45}$ for consecutive pairs of items). A higher discrimination index gave more weight to the interaction term. The discrimination indices for the other terms, $\theta - \beta_i$, were 1 in all cases. In sum, there were $3 \times 5 \times 2 = 30$ conditions; 50 data matrices were generated within each condition.

Models Estimated

Five models were fit for all 1,500 simulated data matrices:

1. M1, a Rasch model,
2. M2, a COIM with interactions between all adjacent items and varying interaction parameters (called "different constant interactions"),
3. M3, a COIM with interactions between all adjacent items and interaction parameters constrained to be equal (called "equal constant interactions"),
4. M4, a DDIM with equal interactions between adjacent items and $a_{ij} = 1$, and
5. M5, a DDIM with equal interactions between adjacent items but a_{ij} equal to its true value (a "true" model).

The parameters of these models were estimated through loglinear analyses. It can be shown that for the models in Equations 1–4, sufficient statistics exist for the θs (the likelihood functions for the models are members of the exponential family, e.g., Andersen, 1980). Therefore, the probability of a response pattern given an appropriate sufficient statistic depends only on the item parameters and not on θ. With these parameterized probabilities, it is possible to design a loglinear model.

Kelderman (1984) did this for the Rasch model as well as for other interaction models. Similar expressions for the models in Equations 2–4 can be derived. Kelderman (1984) showed that the estimates for the item parameters from the loglinear model are equal to the estimates that would be obtained by maximizing the conditional likelihood (Andersen, 1980). This is also true for the more general interaction models, as long as the discrimination indices are not parameters that must be estimated.

The models were estimated using loglinear modeling with SPSS (Norušis, 1994). Ten Vergert, Gillespie, & Kingma (1993) described how this can be done for the Rasch model. Hoskens & De Boeck (1997) also provided some useful suggestions for estimating the proposed models, e.g., using MULTILOG (Thissen, 1988), which allows the user to estimate discrimination parameters for individual items and interaction terms.

Test Statistics

By applying loglinear models, it is also possible to apply the fit criteria used in standard loglinear theory (Agresti, 1990; Bishop, Fienberg, & Holland, 1975). Suppose that n_i is the observed frequency in cell i of the contingency table with N cells and $\hat{m}_i$ is the expected frequency in cell i, based on maximum likelihood estimation of the model parameters. The likelihood ratio χ^2 (G^2) was defined as

$$G^2 = 2\sum_{i=1}^{N} n_i \ln\left(\frac{n_i}{\hat{m}_i}\right) . \tag{5}$$

As $n_i \to \infty$, G^2 is χ^2 distributed with DF degrees of freedom, where DF is equal to the number of free cells minus the number of linearly independent parameters. A second statistic is a modification of G^2:

$$F = \frac{G^2}{DF} , \tag{6}$$

where G^2 is the likelihood ratio χ^2 with DF degrees of freedom. Under the null hypothesis (the model fits), F should have an expected value of approximately 1. F is the deviation of the model from the data, corrected for DF. It allows for comparison between models that have different DFs as a result of either estimation problems or a difference in DF.

These two test statistics were used as follows. First, the number of times that the likelihood ratio χ^2 reached significance at $\alpha = .05$ was counted for all 50 replicated datasets within every condition. Second, the mean F was plotted as a function of the levels of the simulation design variables.

Results

Frequency of Significance of G^2 at $\alpha = .05$

As the results in Table 1 show, the frequency of rejection at $\alpha = .05$ increased with n for the incorrect models. The rejection rate was higher because the power increased with n. The expected number of rejections for the true model over all conditions with the same sample size was 25 $(.05 \times 500)$. The number of actual rejections of the true model was 8 (for $n = 100$), 36 (for $n = 500$) and 24 (for $n = 5,000$). The difference between expected and observed rejection frequencies was significant [$\chi^2(2) = 16.44$, $p < .005$]. This significant effect was primarily due to the very low number of significant tests for $n = 100$. In that case, G^2 was not χ^2 distributed because of the small number of observations in each cell.

Table 1
Frequency of Significance of G^2 at $\alpha = .05$ for M1–M5 and Combinations of n, β_{ij}, and a_{ij}

n	β_{ij}	a_{ij}	M1	M2	M3	M4	M5
100	−2.0	1	8	2	0	0	0
100	−2.0	2	4	0	2	0	0
100	−.5	1	5	8	7	1	1
100	−.5	2	12	10	7	0	0
100	0.0	1	12	6	10	4	4
100	0.0	2	28	12	13	4	1
100	.5	1	28	4	5	1	1
100	.5	2	50	10	10	1	0
100	2.0	1	50	0	2	1	1
100	2.0	2	50	0	2	0	0
500	−2.0	1	50	6	10	1	1
500	−2.0	2	48	1	4	0	1
500	−.5	1	32	28	31	4	4
500	−.5	2	50	48	50	7	4
500	0.0	1	46	35	38	6	6
500	0.0	2	50	49	50	13	6
500	.5	1	50	41	44	7	7
500	.5	2	50	49	50	13	4
500	2.0	1	50	27	35	3	3
500	2.0	2	50	32	39	0	0
5,000	−2.0	1	50	49	50	4	4
5,000	−2.0	2	50	49	50	15	3
5,000	−.5	1	50	50	50	3	3
5,000	−.5	2	50	50	50	50	2
5,000	0.0	1	50	50	50	1	1
5,000	0.0	2	50	50	50	50	2
5,000	.5	1	50	50	50	4	4
5,000	.5	2	50	50	50	50	2
5,000	2.0	1	50	50	50	3	3
5,000	2.0	2	50	50	50	38	0

The more restrictive models (M2 and M3) had more rejections of the null hypothesis than M4, but M3 did not have many more rejections than M2. This can be attributed to the underlying structure of the data, which had equal interactions between adjacent items. M1 had the poorest fit to the data.

With respect to the weight of the interaction term, it was found that $a_{ij} = 2$ resulted in more rejections of the COIM (unless $n = 100$ and β_{ij} was extreme). However, finding the correct weight of the interaction term required a sample size larger than 500. Note that when $a_{ij} = 1$, the number of rejections for M4 was equal to that for M5 because the models for those conditions were equal.

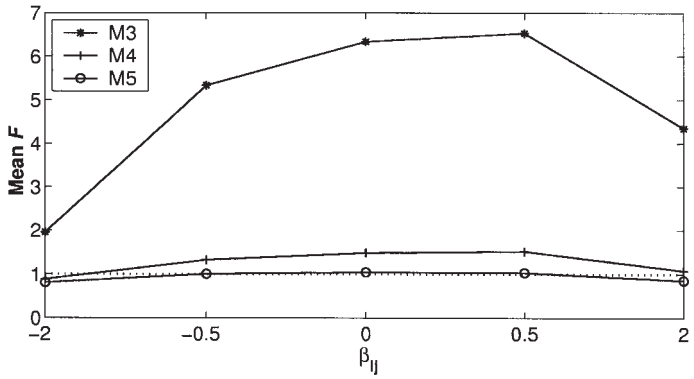
Mean *F* Values

For the sample size and discrimination index variables, mean F results agreed with the G^2 results in Table 1. Mean F values increased with n for the incorrect models, but not for the true model (M5). Also, mean Fs were higher when the discrimination index was 2 compared to when it was 1.

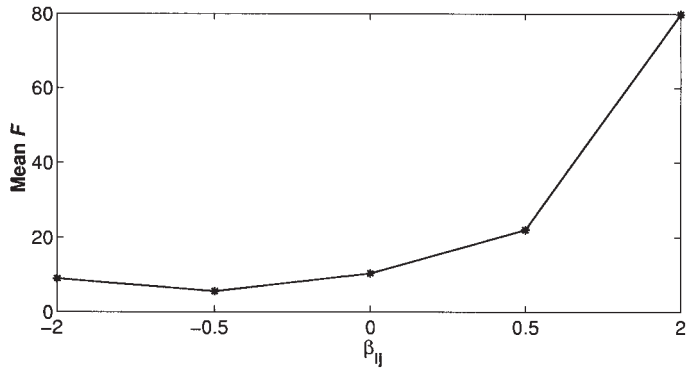
For the interaction parameter, mean F had a clear asymmetry (results for M3, M4 and M5 are in Figure 2a; results for M1 are in Figure 2b because of the magnitude of the Fs; results for M2 are not shown because they were almost equal to those of M3).

Figure 2
Main Effects, Collapsed Across Levels of n and a_{ij}, for Mean F as a Function of β_{ij}

a. M3, M4 and M5



b. M1



The asymmetric relation shown in Figure 2 was very strong for M3. This relation was due to the very large effect for $n = 5{,}000$, but there was also an asymmetric relation for $n = 100$ and $n = 500$. Also, for M3, M4, and M5 (Figure 2a) the mean F, similar to the individual Fs, reached its maximum for $\beta_{ij} = .5$ and not for $\beta_{ij} = 0.0$. For M1 (Figure 2b), the pattern was different than for M3, M4, and M5: higher misfit was associated with higher levels of β_{ij}.

An Unexpected Result

It was expected that the datasets generated with $\beta_{ij} = 0$ would deviate more from the COIM than datasets generated with other values of β_{ij}. In the latter, there is always a majority of examinees with a positive or a negative interaction, which can be captured to some degree with a COIM. Samples with other values of β_{ij} would deviate less from a constant interaction sample. As a result, the fit of a COIM was expected to be better for these samples than for a sample in which $\beta_{ij} = 0.0$. This reasoning seems to explain why a COIM fits better for extreme values of β_{ij}, but it cannot explain the asymmetry of the evaluation criteria around $\beta_{ij} = 0$ for M3, M4 and M5, shown in Figure 2a.

A reasonable explanation for this result might be as follows. Two aspects of the manifest data are of importance: the consistency within response patterns, and the heterogeneity between response patterns in the sample with respect to the direction of the interaction. When β_{ij} is very negative, almost all examinees will exhibit a consistent answer pattern (all 1s), and the direction of interaction will have a high degree of homogeneity. But when the interaction parameter increases, the inconsistency within the answer patterns as well as heterogeneity with respect to the direction of the interaction increase. Interaction heterogeneity reaches its peak at the point where the interaction parameter is at the mode of the θ distribution. From that point, the interaction heterogeneity begins to decrease with increasing β_{ij}. But the inconsistency within the response patterns continues to increase with increasing β_{ij}. Hence, the COIM has the most difficulty in fitting the data when β_{ij} is at an optimal combination point of heterogeneity and inconsistency, and is beyond $\beta_{ij} = 0$.

For the relation between mean F and the level of interaction for M1 (Figure 2b), two points should be stressed. First, there was an overall increase in the level of misfit for M1 because the Rasch model has no parameter that can account for interaction in the data. Second, if the interaction parameter becomes positive, and hence the interaction becomes negative, more inconsistent response patterns in the sample occur and the fit of the Rasch model to the data further decreases. The Rasch model has more difficulty fitting data that are inconsistent.

Conclusions

Three conclusions can be drawn concerning the possibility of distinguishing between the processes of constant and dimension-dependent interaction. First, when the interaction parameter β_{ij} is not too large, dimension-dependent interaction is statistically detectable with 500 examinees. It is also sometimes detectable with 100 examinees. Second, a larger discrimination index for the interaction term increases the rejection rate for the COIM. However, detecting the true value of the weight index a_{ij} requires a sample size of at least $n = 500$; if β_{ij} is extreme, larger n might be required. Third, the optimal value of the dimension-dependent interaction parameter for a maximal distinction between constant and dimension-dependent interaction is .5, which implies that there is an asymmetric relation between the value of the interaction parameter and the capacity to reject an incorrect COIM.

Although these conclusions are based on a simulation that used only five items, there is no reason to expect different results with a larger number of items. Thus, if there is good reason to suspect a dimension-dependent interaction, as when learning occurs during a test, it is possible to detect it, provided that the sample size is at least $n = 500$.

References

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

Andersen, E. B. (1980). *Discrete statistical models with social sciences applications.* Amsterdam: North-Holland.

Bell, R. C., Pattison, P. E., & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement, 12,* 15–26.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge MA: MIT Press.

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.

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

Norušis, M. J. (1994). *SPSS advanced statistics* (Version 6.1) [Computer software]. Chicago: SPSS Inc.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. *Communications of the ACM, 32,* 1192–1201.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Ex-

panded edition (1980), with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.

Ten Vergert, E., Gillespie, M., & Kingma, J. (1993). Testing the assumptions and interpreting the results of the Rasch model using loglinear procedures in SPSS. *Behavioral Research Methods, Instruments, & Computers, 25,* 350–359.

Thissen, D. (1988). *MULTILOG* [Computer software]. Mooresville IN: Scientific Software.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wilson, M., & Adams, R. A. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181–198.

Acknowledgments

The first author is a Research Assistant of the Fund for Scientific Research, Flanders, Belgium.

Author's Address

Send requests for reprints or further information to Francis Tuerlinckx, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Francis.Tuerlinckx@psy.kuleuven.ac.be.