Multivariate Behavioral Research, 34 (2), 245-268
Copyright © 1999, Lawrence Erlbaum Associates, Inc.

Confirmatory Analyses of Componential Test Structure Using Multidimensional Item Response Theory

Rianne Janssen and Paul De Boeck
University of Leuven, Belgium

The componential structure of synonym tasks is investigated using confirmatory multidimensional two-parameter IRT models. It was hypothesized that an open synonym task is decomposable into generating synonym candidates and evaluating these candidate words with respect to their synonymy with the stimulus word. Two subtasks were constructed to identify these two components. Different confirmatory models were estimated both with TESTMAP and with NOHARM. The componential hypothesis was supported, but it was found that the generation subtask also involved some evaluation and that generation and evaluation were highly correlated.

Parallel with the advent of a cognitive stream in test theory (see, e.g., Frederiksen, Mislevy, & Bejar, 1993; Mislevy, 1996), extensions of the standard models in Item Response Theory (IRT) have been proposed, in which parameters are included to model more elementary, cognitive psychological aspects or cognitive components of item solving. Examples of such componential IRT models can be found in Bock, Gibbons, and Muraki (1988), Butter, De Boeck, and Verhelst (1998), Embretson (1980, 1984), Fischer (1973, 1983), Hoskens and De Boeck (1995), Jannarone (1986), Kelderman and Rijkes (1994), Maris (1995), Stegelmann (1983), and van Leeuwe and Roskam (1991). The cognitive components can refer to cognitive processes, to multiple latent abilities involved in item solving, or to features of item difficulty. The IRT modeling of these components is done within either a unidimensional or a multidimensional latent space, making use of dichotomous or polytomous item scores.

The research was carried out while the first author was a research assistant of the National Fund for Scientific Research of Belgium. The manuscript was written while she was working at the Methodology Department of the Faculty of Social Sciences, University of Tilburg, the Netherlands. The revision of the manuscript was done as a postdoctoral researcher funded by the University of Leuven (PDM 97/042). The research was also supported by a research grant (OT 89/9) from the University of Leuven. We wish to thank Luc Delbeke, Iven Van Mechelen and the anonymous reviewers for their comments on previous versions of the manuscript.

Address all correspondence to Paul De Boeck, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. E-mail: Paul.DeBoeck@psy.kuleuven.ac.be

According to Embretson (1983) an important advantage of componential IRT models as a general methodology for assessing the componential structure of a task is that these models measure both the items and the persons on the constructs involved in the processing of a task. Moreover, they specify how item and person characteristics interact to produce response potential. In comparison with a cognitive, experimental approach, componential IRT models do also provide an individual differences model beside a model for cognitive processes, whereas in an experimental approach individual differences are not dealt with in an explicit way. In comparison with a factor-analytic approach, componential IRT models do also provide a model for how characteristics of the task and items affect the responses, whereas in factor analysis components are only based on correlations of individual differences between tasks.

Componential IRT modeling needs a carefully constructed test design (Embretson, 1985), involving a prior conceptualization of the cognitive structure of the items and of how the hypothesized structure can be assessed by means of the selected IRT model. A test design is similar to an experimental design. The cognitive components are determined either by the manipulation of the item characteristics of the selected items, or by the manipulation of the tasks that have to be performed by the subjects with the item stems. These two possibilities of manipulation for a componential item design are equivalent to what Embretson (1983) described as the method of complexity factors and the method of subtask responses, respectively. The *method of complexity factors* is strongly related to the linear logistic latent trait model (LLTM) of Fischer (1973, 1983), where the item difficulty parameter of the Rasch model is decomposed into a linear function of a common set of basic item difficulty factors, describing the item difficulty with respect to the cognitive operations needed in solving the item. The identification of the underlying cognitive principles of item difficulty is only possible in an item set that is constructed on the basis of these (hypothesized) principles. In the *method of subtask responses*, subjects have to perform a series of subtasks with the same item stem. Cognitive components intervening in the solution process of the total item are identified from the responses to these subtasks. A verbal analogy task, for example, can be decomposed into a rule construction and a response evaluation subtask (Embretson, 1980), both designed to assess the corresponding covert cognitive components involved in the total task of solving an analogy. Note that the difference between the method of complexity factors and the method of subtask responses has implications for the task instructions for the subjects, which are the same for all items in the method of complexity factors and different in the method of subtask responses.

Componential IRT models can be classified on the basis of whether they need data gathered according to the subtask method or whether a design of the items in terms of their complexities suffices. For example, for the application of the multicomponent latent trait model (Embretson, 1980, 1984) or the conjunctive Rasch model (Maris, 1995) subtask data are needed, while for the multicomponent Rasch model (Stegelmann, 1983) or the conjunctive IRT model (van Leeuwe & Roskam, 1991) a design of item complexity suffices. Janssen and De Boeck (1997) described two other general classification criteria for componential IRT models: (a) the type of linkage rule, and (b) the weighting of the components. The *linkage rule* describes the relationship between the components and the performance in the total task. In compensatory models, a high ability on one component can compensate for low abilities on other components in solving a task. Conjunctive models on the other hand imply that a minimum competence is needed in each component for solving a task. Examples of compensatory models are the multidimensional two-parameter logistic model (Reckase & McKinley, 1982) and the multicomponent Rasch model (Stegelmann, 1983). Conjunctive IRT models are described by Embretson (1980, 1984), Maris (1995), and van Leeuwe and Roskam (1991). The *weighting of the components* refers to whether the components are equally weighted within each item and over all items or whether a componential weight parameter is included to measure the differential dependence of the composite item on the components. Componential IRT models that are extensions of the Rasch model (like, e.g., Maris, 1995; Stegelmann, 1983) do not have weight parameters. For the moment, linkage models with weight paramaters for each item can only be found among the compensatory IRT models.

Janssen and De Boeck (1997) showed on the basis of a heuristic evaluation procedure that for the psychometric modeling of componentially designed synonym tasks, models with item-specific componential weights were clearly better than models with unweighted linkage. Compensatory and conjunctive models had approximately equally good fit. Therefore, it is expected that a confirmatory version of the multidimensional two-parameter IRT model would result in good fit for the synonym data. It is the purpose of the present article to analyze the data of Janssen and De Boeck with this model. In the following, the multidimensional two-parameter IRT model and its confirmatory version are discussed first. Next, it is discussed how different confirmatory models can be compared with respect to their goodness of fit. Finally, the componential hypothesis of the synonym tasks is described and different confirmatory models for the data are presented.

The Multidimensional Two-parameter IRT Model

In the (exploratory) multidimensional two-parameter (M2P) IRT model a person *j* (*j* = 1, ..., *J*) is characterized by an ability parameter θ_{jk} on each of *K* dimensions. An item *i* (*i* = 1, ..., *I*) is characterized by a global item difficulty β_i and a vector $\boldsymbol{\sigma}_i$ of *K* discrimination parameters. The probability of success on an item *i* by a person *j* is function of the difference between a weighted sum of abilities and the item difficulty:

$$(1) \qquad P(X_{ij} = 1; \theta_j, \boldsymbol{\sigma}_i, \beta_i) = f\left(\sum_{k=1}^{K} \sigma_{ik}\theta_{jk} - \beta_i\right).$$

The model assumes that depending on the item, a different weighted sum of abilities is invoked to compete with item difficulty. In a strict extension of the unidimensional two-parameter IRT model, one would expect β_i to be equal to $\Sigma\sigma_{ik}\beta_{ik}$, but the item difficulty parameters cannot be identified on the separate dimensions, but only on the scale of the weighted sum of all person abilities, which is different from item to item. The item discrimination parameters σ_{ik} indicate the slope of the logistic regression lines relating the k^{th} latent ability dimension to the success probability. These parameters correspond to the sensitivity of the item to the k^{th} ability dimension and are equivalent to the factor loadings of an item in the traditional factor model.

The function *f* in Equation 1 can either be the logistic function, resulting in the M2P logistic (M2P-L) model (McKinley & Reckase, 1983; Reckase & McKinley, 1982), or the cumulative normal distribution, resulting in the M2P normal ogive (M2P-NO) model (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). The M2P-NO model is often called full-information item factor analysis, as Takane and de Leeuw (1987) showed that the marginal likelihood of the model is formally equivalent with the likelihood of the factor analysis models for binary data (see, e.g., Muthén, 1984).

Confirmatory Versions of the M2P Model

When the M2P model is used in componential research, a confirmatory version of it is needed in order be able to specify the hypothesized componential test structure. In such a confirmatory version, Equation 1 is supplemented with an item structure vector $\mathbf{s}_i$ of order *K*:

$$(2) \qquad P(X_{ij} = 1 | \mathbf{s}_i; \theta_j, \boldsymbol{\sigma}_i, \beta_i) = f\left(\sum_{k=1}^{K} \sigma_{ik} s_{ik} \theta_{jk} - \beta_i\right).$$

Each s_{ik} is an indicator variable with a value of 1 indicating that the item i is measuring the k^{th} ability dimension and a value of 0 indicating that it is not. Hence, the componential design of the test can be represented in the design of the item discrimination parameters by constraining a subset of item loadings to be zero. By consequence, the confirmatory M2P model specifies the componential structure of the items with respect to the person abilities involved. When the componential data are gathered according to the method of subtasks, the difference between a total task item and its subtask items can be modeled with the restriction that the separate dimensions involved in the subtask items of a given type also intervene in the corresponding total task item, but not in the subtask items of another type.

For the confirmatory M2P-L model, McKinley (1988, 1989, 1992; see also McKinley & Kingston, 1988) developed the program TESTMAP. This program uses the Marginal Maximum Likelihood estimation procedure supplemented with the EM algorithm (Bock & Aitkin, 1981). A multivariate normal distribution is assumed for the person ability parameters with the means fixed to zero and the variance-covariance matrix of the ability parameters equal to the identity matrix. Confirmatory M2P-L models can also be specified within the general framework of Multidimensional Polytomous Latent Trait (MPLT) models of Kelderman and Rijkes (1994), within which models for dichotomous items form a subclass. However, characteristic for the MPLT models is that the dependence of the item responses on the latent traits has to be specified for all items by choosing for each item i a priori scoring weights w_{ik} on each dimension k. The possible values of w_{ik} are restricted to integers. The advantage of this kind of weights is that the resulting models are all a member of the exponential family, and that consequently sufficient statistics exist for the person parameters, so that conditional maximum likelihood estimation for the item parameters is possible. Unlike in the MPLT models, the values of the weights in the confirmatory M2P model do not need to be specified in advance.

For the M2P-NO model, the program NOHARM of Fraser (1988) allows to estimate both exploratory and confirmatory models. The estimation method of the program makes use of bivariate information only (i.e., item means and covariances). In contrast with NOHARM, the program TESTFACT (Wilson, Wood, & Gibbons, 1984) uses the full information in the response patterns for the estimation of the item parameters, but it can only be used for exploratory analysis. Furthermore, Gibbons and Hedeker (1992) developed full-information item bi-factor analysis, which consists of a K-dimensional solution, with one general ability and K - 1 group or method related ability dimensions. The bi-factor structure constrains each item to load on the general factor and on only one of the K - 1 group factors. Hence, the model of Gibbons and Hedeker can be seen as a special case of the confirmatory M2P-NO model, as full-information item bi-factor analysis allows only for special item structure vectors, namely of the bi-factor type.

For the analysis of the componentially designed synonym tasks, a versatile program was needed in order to be able to estimate different confirmatory M2P models. Hence, only TESTMAP and NOHARM remained as possible candidates. Both programs differ with respect to the estimation procedure used, but, as a consequence of this, they also differ with respect to the way the goodness of fit of different models are compared. It will be explained in the following that the latter fact motivated us to estimate the confirmatory M2P models for the synonym data with both programs. In the next section, the two approaches to goodness-of-fit testing are described first. Afterwards, a comparison is given.

Model Selection

In TESTMAP, the goodness-of-fit of different models are compared on the basis of two criteria originally derived within the framework of information theory, namely Akaike's Information Criterion (AIC; see, e.g., Akaike, 1977) and its Consistent version (CAIC; Bozdogan, 1987). Both model selection criteria choose the best approximating model among a set of competing models for a given data set taking model complexity into account. In contrast with likelihood ratio tests, the competing models do not need to be nested. The AIC equals

$$\text{AIC} = -2ln(L) + 2m, \tag{3}$$

where $ln(L)$ is the natural logarithm of the likelihood in the maximum likelihood solution and m denotes the number of estimated parameters. The first term of the sum is an index of the distance between the estimated model and the true model: the greater the likelihood of the solution, the closer the fitted model is presumed to approximate the true model, but the lower the negative loglikelihood. The second term of the sum constitutes a penalty for model complexity. Hence, the AIC has to be minimized to choose the optimal (and most parsimonious) model from a set of models. Bozdogan developed an extension of the AIC, called the consistent AIC or CAIC. It was derived in order to make the AIC asymptotically consistent and to penalize overparameterization more stringently. The CAIC equals:

$$\text{CAIC} = -2ln(L) + m[ln(n) + 1], \tag{4}$$

with *n* denoting the sample size. Minimization of the CAIC generally leads to simpler models than those obtained by minimizing the AIC.

In NOHARM, the goodness of fit of different models is based on the residual covariances of the model. McDonald and Mok (1995) adopted the unweighted least squares (ULS) goodness-of-fit index of Tanaka (1993) for factor analysis for multidimensional IRT models:

$$\gamma_{\mathrm{ULS}} = 1 - [\mathrm{Tr}(\mathbf{R}^2)]/[\mathrm{Tr}(\mathbf{C}^2)], \tag{5}$$

where **C** is the sample item covariance matrix and **R** the item residual covariance matrix. γ_{ULS} is a descriptive measure of goodness of fit, indicating how much of the item covariances is explained by the model. As it does not take model complexity into account, one may expect the measure to be related to the number of parameters in the model.

McDonald and Marsh (1990) warned against the use of the AIC as a way of testing a model in the context of structural equation models. The value of the AIC would depend on sample size, like that of the conventional chi-square test. As a consequence the AIC would tend to prefer saturated models in very large samples and models with few estimated parameters in very small samples. Nevertheless, McKinley (1989) successfully applied the AIC and CAIC in a small simulation study with confirmatory M2P-L models. McDonald and Mok (1995) argued that further simulation studies involving the resampling of real data at different sample sizes are needed to resolve the issue of the quality of the (C)AIC for multidimensional IRT models. To make sure about our results, we decided to use both the two information criteria and the γ_{ULS} statistic. The AIC and CAIC are used as measures of relative goodness of fit among a set of competing models. The γ_{ULS} statistic is used as a measure of the absolute degree of approximation of the model to the data. As the goodness-of-fit measures to be used are based on two different estimation methods (and computer programs), the correspondence between the estimated parameters will be investigated as well.

The Componential Structure of Synonym Tasks

Previous studies (Butter, De Boeck, & Baele, 1992; Janssen, De Boeck, & Vander Steene, 1996; Janssen, Hoskens, & De Boeck, 1993) have investigated the componential structure of an *open synonym task* (i.e., a task in which a synonym must be provided for a stimulus word). The open synonym task was considered a total task, this is a composite task to be decomposed into subtasks. It was hypothesized that solving an item of the open synonym task is based on the generation of synonym candidates and the subsequent or concurrent evaluation of these synonym candidates on their synonymy with the stimulus word. Two subtasks were designed to identify the two cognitive components of the open synonym task. In the *generation subtask*, respondents listed all words that came to their mind while searching for a synonym. In the *evaluation subtask*, the respondents had to select the "true" synonym(s) of a stimulus word from a list of four words. The three distractors were selected from the most frequently provided words in the generation subtask. Table 1 contains an example of the open synonym task and the two subtasks for the stimulus word "foggy".

On the basis of structural equation modeling using the sum scores in the three tasks, Janssen et al. (1996) showed that the generation component ability is primarily related to verbal fluency abilities, while the evaluation component ability is primarily related to verbal comprehension abilities. Moreover, evidence was provided that these two component abilities could account for the correlations of the open synonym task with other ability measures. In an analysis based on the proportion of correct responses on the items, Butter, De Boeck, & Baele (1992) provided some evidence for the differential validity of the subtask difficulties. The items of the evaluation subtask were less difficult, the more similar the stimulus word was to the correct response, according to ratings of similarity in meaning, use, and associations. The item difficulty of the generation subtask was also determined by this similarity factor. However, in contrast with the evaluation subtask, the item difficulty for the generation subtask was also determined by a word availability factor, which summarized the rated power of the stimulus word to evoke contexts, images, and associations. Highly available stimulus words were less difficult for generating a synonym.

Table 1
Example of the Three Synonym Tasks for the Stimulus Word "Foggy"

Task	Example
Generation	Which words come to your mind while searching for a synonym for "foggy"?
Evaluation	Which of these words do you consider to be a synonym for "foggy"? a) damp b) cloudy c) blurred d) hazy
Open Synonym	Give a true synonym for "foggy".

Note. The original tasks were presented in Dutch.

Confirmatory Models for the Synonym Tasks

In the present article, the componential hypothesis will be studied making use of the confirmatory M2P model. The item structure for different componential models are presented in Table 2. The models will be described first. Afterwards, an interpretation is given.

In the initial confirmatory model, it is hypothesized that the items of the generation and evaluation task each measure a separate ability dimension, and that both ability dimensions are involved in the open synonym task. This initial confirmatory model is labeled *model 2D_ge* in Table 2, as it is a two-dimensional model that restricts the items of the generation and of the evaluation task each to one dimension. Model 2D_ge assumes that the two ability dimensions are uncorrelated. This assumption of orthogonal ability dimensions is always made in TESTMAP, as it assumes a multivariate normal distribution for the ability parameters with variance-covariance matrix equal to the identity matrix for the MML-procedure. However, when generation and evaluation are in fact correlated, the solution of model 2D_ge cannot identify these two dimensions correctly. NOHARM on the other hand offers the possibility to estimate models with orthogonal or with correlated ability dimensions. The estimation of a confirmatory model with correlated ability dimensions for the item structure vectors of model 2D_ge will be labeled as *model 2D_ge(r)*. Models 2D_ge and 2D_ge(r) estimated with NOHARM, can be compared for their goodness of fit to decide upon generation and evaluation being correlated or not.

For comparative reasons, several other componential models were formulated. First, also the unrestricted one-, two-, and three-dimensional M2P models were estimated. These models are labeled as *model 1D*, *model 2D*, and *model 3D*, respectively. They correspond to confirmatory M2P models with for all items a unit vector of order K for $\mathbf{s}_i$ (yielding in fact an exploratory analysis). Second, two other confirmatory two-dimensional

Table 2
Item Structure Vectors for the Confirmatory Models for the Synonym Tasks

	Model						
Task	2D_ge	1D	2D	3D	2D_e	2D_g	3D_ge
Generation	1 0	1	1 1	1 1 1	1 1	1 0	1 0 1
Evaluation	0 1	1	1 1	1 1 1	0 1	1 1	0 1 1
Open Synonym	1 1	1	1 1	1 1 1	1 1	1 1	1 1 1

models were estimated. In *model 2D_e*, the items of the *evaluation* task are restricted to measure only one dimension, whereas the items of the generation task and of the open synonym task are assumed to measure both dimensions. This model guarantees that one dimension coincides with the evaluation items. *Model 2D_g* is the complement of model 2D_e in that it restricts the items of the *generation* task to be unidimensional, but not the items of the other two tasks. Items from the evaluation task and from the open synonym task are supposed to measure both dimensions. This model guarantees that one dimension coincides with the generation items. Finally, a three-dimensional confirmatory model was estimated, which is labeled *model 3D_ge*. The first two dimensions comprise the initial two-dimensional model 2D_ge, with a specific generation and evaluation dimension as in model 2D_ge, but with a general third dimension in addition.

The models in Table 2 differ with respect to the dimensionality of the latent space of the synonym tasks, the supposed componential structure of the items, and whether the abilities involved in the generation and evaluation subtask are correlated or not. Note that models 2D_g, 2D_e, and 3D_ge do not only differ from model 2D_ge with respect to their componential structure, but also with respect to whether a correlation is allowed between the abilities involved in the generation and evaluation subtask. Models 2D_e, 2D_g, and 3D_ge do show a componential overlap between the generation and evaluation subtasks, and, as a consequence, the subtasks are assumed to correlate, even within the ortohogonal structure of TESTMAP. Models 2D_e and 2D_g have in common that one type of subtask is related with only one dimension while the complementary subtask can load on two dimensions. For model 2D_e, this results in a correlation over persons between the evaluation ability and the logit of the probability of success in an item *i* of the generation task. This can be derived analytically and from the geometric properties of the model:

$$(6) \qquad r(\theta_{j2}, \sigma_{i1}\theta_{j1} + \sigma_{i2}\theta_{j2} - \beta_i) = r(\theta_{j2}, \sigma_{i1}\theta_{j1} + \sigma_{i2}\theta_{j2})$$

$$= \frac{\sigma_{i2}}{\sqrt{\sigma_{i1}^2 + \sigma_{i2}^2}} = \frac{\sigma_{i2}}{\|\boldsymbol{\sigma}_i\|}$$

$$= \cos(\alpha_i)$$

with $\|\boldsymbol{\sigma}_i\|$ being the vector length of $\boldsymbol{\sigma}_i$ and α_i being the angle of $\boldsymbol{\sigma}_i$ with the evaluation ability axis. The calculation is based on the assumption that the variance-covariance matrix of the ability parameter vector is the identity matrix. This assumption is always made in TESTMAP and can be made in NOHARM when an orthogonal solution is chosen. For model 2D_g, the correlation over persons between the generation ability and the logit of the probability of success in an item of the evaluation task can be derived similarly. This correlation equals the cosine of the angle of $\boldsymbol{\sigma}_i$ of an item of the evaluation task and the generation ability axis.

Note that in Equation 6 the correlation is estimated on the basis of one item. An estimate of the correlation over all the items can be obtained by taking the cosine of the mean of the angles α_i. We calculated a weigthed mean angle over the items with as weights the vector length $\|\boldsymbol{\sigma}_i\|$ of the items of the subtask involved:

$$\bar{\alpha} = \frac{\sum_{i=1}^{I} (\|\boldsymbol{\sigma}_i\| \alpha_i)}{\sum_{i=1}^{I} \|\boldsymbol{\sigma}_i\|} \tag{7}$$

with α_i being the angle of $\boldsymbol{\sigma}_i$ of an item *i* with the reference axis. The weighting is done in order to diminish the influence of points near the origin, where variation in the angle between $\boldsymbol{\sigma}_i$ and the reference axis is less meaningful.

In Model 3D_ge, the general factor can account for the intercorrelations between the two types of subtasks. It can be shown that the correlation over persons between the logit of the probability of success in an item *i* of the generation task and the logit of the probability of success in an item *i′* of the evaluation task equals the product of the correlations of the ability of the third dimension with the logit of item *i* of the generation task and with the logit of item *i′* of the evaluation task:

$$\begin{aligned}
(8)\quad & r(\sigma_{i1}\theta_{j1} + \sigma_{i3}\theta_{j3} - \beta_i, \sigma_{i'2}\theta_{j2} + \sigma_{i'3}\theta_{j3} - \beta_{i'}) \\
&= \frac{\sigma_{i3}\sigma_{i'3}}{\|\boldsymbol{\sigma}_i\|\|\boldsymbol{\sigma}_{i'}\|} \\
&= \cos(\alpha_i)\cos(\alpha_{i'}) \\
&= r(\theta_{j3}, \sigma_{i1}\theta_{j1} + \sigma_{i3}\theta_{j3}) r(\theta_{j3}, \sigma_{i'2}\theta_{j2} + \sigma_{i'3}\theta_{j3})
\end{aligned}$$

with α_i (and $\alpha_{i'}$) now being the angle of $\boldsymbol{\sigma}_i$ (or $\boldsymbol{\sigma}_{i'}$) with the third axis. Calculating the latter product using the cosine of the weighted mean angles for both subtasks (cf. Equation 7) gives an estimate over all items of the correlation between the abilities involved in the generation subtask and in the evaluation subtask.

Note that it is not necessary to estimate models 2D_g, 2D_e, and 3D_ge with correlated ability dimensions in NOHARM (like with model 2D_ge(r) for model 2D_ge), as these models already allow for a correlation between the two subtasks. In fact, the goodness of fit of these models and their counterparts with correlated ability dimensions as estimated with NOHARM were exactly the same.

Method

As the present article reanalyzes the data of Janssen and De Boeck (1997), this method section only summarizes the most relevant aspects of the method section in their article.

Items

Janssen and De Boeck (1997) compiled a list of 120 Dutch stimulus words paired with a synonym according to the dictionary. The list of 120 items was divided at random into two lists of 66 word pairs, with an overlap of 12 items. The two lists and their overlap each consisted of an equal number of nouns, verbs, and adjectives. The stimulus words of the three grammatical word classes were grouped on separate sheets of paper. This grammatical classification of items was carried out in order to avoid certain ambiguities in Dutch where some words can refer to a verb as well as to the plural of a noun (e.g., “dingen” meaning both “to compete for” and “things”). The order of presentation of word classes was randomized across subjects and tasks. The order of the stimulus words within a word class remained the same in the three tasks used.

Respondents

Pupils from the last two grades from six different Dutch-speaking Belgian schools of general secondary education participated during school time. Their ages varied between 16 and 18. Of the subjects answering the items of the first list, 218 completed all three tasks. For those working with the items of the second list, this number was 258. Seven subjects were discarded as they had not followed the instructions of the evaluation task. Hence, the final number of subjects was 212 for List 1, and 257 for List 2.

Procedure

All subjects completed the generation task first. After about six weeks, the open synonym task was administered. The evaluation task was presented in a third session about six weeks after the open synonym task, but due to practical circumstances, about half of the subjects completed the evaluation task in the second session, right after the open synonym task. The time intervals between the tasks were needed to control as much as possible for memory effects between the tasks, and to allow for the construction of the evaluation task (as the response alternatives were based on the responses given to the generation task). Memory effects are especially to be feared among the open-ended format tasks, and from a multiple-choice format to an open-ended format. That is why in each case the generation task was separated from the open synonym task by at least six weeks and why the evaluation task always came last.

Results

Preliminary Remarks

Note that the number of respondents is quite small for the estimation of an M2P model. However, in the present article, we were not so much interested in stable parameter estimates of individual items, but in the general pattern of the loadings of the items on the three synonym tasks and in the goodness of fit of the different models. Moreover, the item parameters were estimated for research purposes only, and were not used for the classification of the subjects. Finally, as a way of cross-validition (but also because of practical restrictions on the number of items in the computer programs), the different confirmatory models were estimated on six subsets of the whole data set. The six data sets contain the data for the three synonym tasks for the adjectives, nouns, or verbs from List 1 or List 2, respectively. Each data set contains 66 items (namely 22 stimulus words × 3 synonym tasks) in principle, but the items that were failed by all respondents in at least one task were excluded for the three tasks. For the adjectives of List 1, and the nouns, verbs, and adjectives of List 2, 22 items remained for the three synonym tasks. This number was 21 for the nouns of List 1, and 19 for the verbs of List 1.

Comparison between TESTMAP and NOHARM

The different confirmatory M2P models presented in Table 2 were all estimated with TESTMAP and NOHARM. Apart from the model specifications, NOHARM does not require any specific settings in the program, but the TESTMAP program requires the user to specify an upper bound for the estimates of the discrimination parameters. This maximum was set to the value specified in the manual (being 1.5) for all the analyses. Because of this difference, and because TESTMAP works on the logistic scale and NOHARM on the normal ogive scale, the correspondence between the estimated parameters of each model was expressed as a correlation coefficient calculated over all the items in the data set for the six data sets separately.

Correlations high in the nineties were found between the *item difficulty parameters* for the same models estimated by the two programs, but also across different models, regardless of their dimensionality and of whether the estimation was done with the same or with different programs. As Table 3 shows, the picture was different for the *item discrimination parameters*. Note that for model 2D and 3D, the correlations were calculated not only for the corresponding dimensions, but also for the cross-pairs of the dimensions. These cross-dimenson correlations are not given for the other models, as the constraints in the item structure vectors forced the dimensions to be well identified. For the models with a fixed zero loading for one type of subtask, the subtask items with the zero loadings were left out of the calculation of the correlation coefficient for the discrimination parameters for that dimension. The correlations between the item discrimination parameters estimated by TESTMAP and by NOHARM were quite satisfying for the more restrictive models, that is, for models with restrictions either on the number of dimensions, or on the subtask loadings. The correlations were high for model 1D and for model 2D_ge, but less good for model 2D and the three-dimensional models. While it was still possible to link the dimensions of model 2D between NOHARM and TESTMAP on the basis of their correlations, this was no longer possible for all data sets in model 3D. For example, for the adjectives of List 1, the first dimension of NOHARM seems to be predominantly related to the third dimension of TESTMAP, while the second and third dimension of NOHARM are both mostly related to the first dimension of TESTMAP. Note also that the correlations among the discrimination parameters were lower for model 2D_g than for model 2D_e, in particular for the second dimension in some data sets.

As a conclusion one can state that the correspondence between the item parameter estimates of TESTMAP and NOHARM was reasonably high, especially for models with a restricted number of parameters. The correspondence between the two estimation methods was the more encouraging since data of a relatively small number of subjects were used. Hence, it seems justified to use the goodness-of-fit statistics derived from

Table 3
Correlations Between the Estimated Item Discrimination Parameters from NOHARM and TESTMAP

		Nouns		Verbs		Adjectives	
Model	Dimension[a]	List 1	List 2	List 1	List 2	List 1	List 2
1D	1 - 1	.96	.90	.93	.96	.96	.98
2D	1 - 1	.74	.79	.71	.31	.53	.92
	2 - 2	-.55	.78	.31	-.32	-.09	.57
	1 - 2	.69	.34	.56	.94	.74	.30
	2 - 1	.53	-.41	-.56	.41	.89	-.74
2D_ge	1 - 1	.89	.96	.80	.87	.81	.82
	2 - 2	.88	.90	.81	.94	.86	.95
2D_g	1 - 1	.87	.87	.94	.93	.97	.95
	2 - 2	-.10	.79	.59	.79	.63	-.05
2D_e	1 - 1	.73	.96	.52	.66	.82	.90
	2 - 2	.88	.84	.96	.94	.96	.99
3D	1 - 1	.64	.30	.75	.38	.26	.83
	1 - 2	.71	.24	.47	.48	.40	.44
	1 - 3	.42	.72	.68	.80	.66	.65
	2 - 1	.59	-.57	.20	-.21	.79	-.73
	2 - 2	-.35	.77	.17	.09	.37	.36
	2 - 3	.23	-.18	.43	.05	.45	-.36
	3 - 1	-.31	.56	-.27	.75	.53	.60
	3 - 2	-.09	.47	.49	-.41	-.08	.32
	3 - 3	.70	-.65	.14	-.10	.07	-.15
3D_ge	1 - 1	.75	.93	.25	.63	.85	.85
	2 - 2	.49	.29	.56	.90	-.42	.77
	3 - 3	.85	.79	.90	.91	.92	.97

[a] When two dimensions are mentioned, the first refers to the NOHARM estimate and the second to the TESTMAP estimate.

TESTMAP and NOHARM together to judge the quality of a model. Note that the correspondence found between the two estimation methods supports the conjecture of McDonald and Mok (1995) that "... bivariate analysis of binary data seems likely to provide adequate evidence of dimensionality as compared with full-information methods" (p. 23). In fact, Knol and Berger (1991) showed in a simulation study that with respect to the goodness of recovery of the item parameters in exploratory analyses of multidimensional data, NOHARM performed as well as TESTFACT, which uses a full-information method for estimating exploratory M2P-NO models.

Model Selection

Table 4 presents the model fit according to the AIC and CAIC for the different models for the six data sets as estimated by TESTMAP and the values of the γ_{ULS} as estimated by NOHARM. The models are arranged according to their number of model parameters *m*. Remember that a lower value indicates better fit for the (C)AIC, and a higher value indicates better fit for the γ_{ULS}. The order of the goodness of fit is indicated between parentheses for each statistic and for each data set.

On the basis of the values of the *AIC,* model 2D_ge always came last, with as an exception the nouns of List 2 at rank 5. In the median ranking of the AIC values over the six data sets, model 3D was the best, followed by 3D_ge, 2D_e, 2D, 1D, 2D_g, and 2D_ge. The values of the *CAIC* resulted in an ordering of the models that is very different from the ordering by the AIC, due to their difference in penalty for overparameterization. On the CAIC, the one-dimensional model always came out as the best model. Model 2D_e had a median rank of 2. The median rank of models 2D_ge and 2D_g were 3 and 3.5, respectively. The unconstrained two-dimensional model (2D) and the two models with three dimensions (3D_ge and 3D) were always ranked last in the same order. Note that the values of the CAIC were generally higher for an increasing number of parameters, but model 2D_e was an exception, as it was better (lower CAIC) than model 2D_ge. The γ_{ULS} *statistics* were reasonably high. Moreover, they were about equally high across data sets (although the values for the verbs of List 2 were somewhat lower), and the ordering of the models according to the γ_{ULS} statistics was remarkably similar across data sets. In general, model fit on the basis of the γ_{ULS} was better, the more parameters there were in the model, but model 2D_ge was a notable exception as its fit was worse than for model 1D. Adding a correlation between the latent dimensions of model 2D_ge, hence, resulting in model 2D_ge(r), gave a better fit, especially for the adjectives of List 1. Among the two-dimensional models, model 2D always had the best fit, but model 2D_e was a consistent and close second.

A surprising finding across the three goodness-of-fit measures was that the initial model 2D_ge did not fit well. The AIC gave model 2D_ge the last place in the majority of data sets and the γ_{ULS} statistics indicated that model

Table 4

Model Fit for the Data Sets of List 1 (*N* = 212) and List 2 (*N*= 257)

		List 1				List 2			
Data Set	Model	*m*	AIC	CAIC	γ_{ULS}	*m*	AIC	CAIC	γ_{ULS}
Nouns	1D	126	10686 (5)	11235 (1)	.73 (7)	132	13907 (7)	14507 (1)	.79 (7)
	2D_ge	147	10793 (7)	11434 (3)	.68 (8)	154	13862 (5)	14562 (3)	.78 (8)
	2D_ge(r)	148	-	-	.74 (6)	155	-	-	.80 (6)
	2D_g	168	10745 (6)	11477 (4)	.75 (5)	176	13869 (6)	14670 (4)	.81 (5)
	2D_e	168	10650 (2)	11382 (2)	.77 (4)	176	13737 (3)	14538 (2)	.84 (4)
	2D	189	10672 (3)	11495 (5)	.79 (2)	198	13814 (4)	14714 (5)	.84 (3)
	3D_ge	210	10677 (4)	11592 (6)	.79 (3)	220	13728 (2)	14728 (6)	.85 (2)
	3D	252	10645 (1)	11743 (7)	.82 (1)	264	13635 (1)	14836 (7)	.88 (1)
Verbs	1D	114	11210 (3)	11706 (1)	.74 (7)	132	16599 (5)	17200 (1)	.66 (7)
	2D_ge	133	11259 (7)	11838 (2)	.70 (8)	154	16664 (7)	17365 (2)	.63 (8)
	2D_ge(r)	134	-	-	.74 (6)	155	-	-	.67 (6)
	2D_g	152	11214 (4)	11876 (4)	.76 (5)	176	16565 (3)	17365 (3)	.69 (5)
	2D_e	152	11183 (1)	11845 (3)	.77 (4)	176	16570 (4)	17370 (4)	.69 (4)
	2D	171	11216 (5)	11961 (5)	.78 (3)	198	16622 (6)	17523 (5)	.71 (3)
	3D_ge	190	11205 (2)	12033 (6)	.79 (2)	220	16542 (2)	17543 (6)	.71 (2)
	3D	228	11231 (6)	12224 (7)	.81 (1)	264	16533 (1)	17734 (7)	.75 (1)
Adjectives	1D	132	11970 (5)	12545 (1)	.73 (7)	132	16141 (6)	16741 (1)	.82 (7)
	2D_ge	154	12101 (7)	12771 (4)	.61 (8)	154	16243 (7)	16944 (4)	.79 (8)
	2D_ge(r)	155	-	-	.74 (6)	155	-	-	.82 (6)
	2D_g	176	12002 (6)	12768 (3)	.76 (5)	176	16118 (5)	16919 (3)	.83 (5)
	2D_e	176	11935 (3)	12702 (2)	.77 (4)	176	16038 (3)	16839 (2)	.84 (4)
	2D	198	11913 (1)	12775 (5)	.78 (3)	198	16058 (4)	16958 (5)	.85 (3)
	3D_ge	220	11962 (4)	12921 (6)	.79 (2)	220	16029 (2)	17030 (6)	.86 (2)
	3D	264	11929 (2)	13079 (7)	.81 (1)	264	16026 (1)	17227 (7)	.87 (1)

1D with less parameters was better than model 2D_ge. A plausible explanation for the latter finding is that model 2D_ge does not allow for a correlation between generation and evaluation. With an angle between generation and evaluation that is smaller than 45°, it can be understandable that model 1D had a better fit, since in that model both subtasks are located on one dimension. The need for a correlation between generation and evaluation was also shown in the better fit of model 2D_ge(r) in comparison with model 2D_ge in NOHARM.

As to the issue whether a higher dimensionality is needed, the AIC and CAIC criteria diverged. The three-dimensional models were favored by the AIC statistic, as they mostly appeared in the first ranks, whereas in the ordering by the CAIC, they appeared always on the last positions. The γ_{ULS} statistic for the NOHARM output showed, however, that only the unconstrained model 3D was really better than the two-dimensional models. It was also shown in the previous section that the solution of model 3D was less stable across NOHARM and TESTMAP. Hence, overall a two-dimensional solution seems to be preferred.

Among the two-dimensional models there was one, namely model 2D_e, which was doing well on both the AIC, CAIC, and the γ_{ULS} statistic. Combining the orderings of the AIC and CAIC, model 2D_e even appears to be the best choice for the present data sets. It was better than the three-dimensional models (on the CAIC) and the one-dimensional model (on the AIC). With NOHARM, model 2D_e was the best two-dimensional model with constraints, with about an equal goodness of fit as model 2D.

Model 2D_e differs from the initial model 2D_ge with respect to the supposed componential structure. In contrast with model 2D_ge, the one latent dimension measured by the evaluation task also intervenes in the generation subtask in model 2D_e, meaning that generation is not pure generation but also involves evaluation. As a consequence, model 2D_e allows for a correlation between the generation and evaluation subtasks. The importance of the componential structure can be seen in the fact that model 2D_g generally showed worse fit than model 2D_e, especially on the γ_{ULS} statistic. It was also shown in the previous section that model 2D_g seemed to result in a less stable solution. An explanation for the difference in fit between model 2D_e and 2D_g can be found in that in model 2D_e, the evaluation task is defined as a subset of the generation task with respect to the abilities (dimensions) involved. This is more plausible than the generation subtask being a subset of the evaluation task (as in model 2D_g), as respondents want their responses to be of some quality. On the other hand, evaluating what is given, like in the evaluation subtask, does not require any generation activity. Note also that model 2D_e is in accordance with a response-format effect, as the open synonym task and the generation task are alike in response format and with respect to their item structure vectors, and both tasks differ on these aspects from the evaluation task.

A Further Look at Model 2D_e

Item Discrimination Parameters

Table 5 presents the averages over the six data sets of the means and standard deviations of the obtained item discrimination parameters for model 2D_e as obtained from TESTMAP. The general pattern of these results was quite comparable with the results obtained from NOHARM. The *mean* discrimination parameters showed relatively comparable values on the two subtasks and on the open synonym task. However, the contribution of the generation dimension in comparison with the evaluation dimension was higher for the generation subtask (.31 and .38) than for the open synonym task (.26 and .39). Hence, the difference between the generation and open synonym task consists in the importance of the generation dimension. These findings are important as the values of the discrimination parameter that are not constrained to zero are estimated freely, so that one still has to look at the values in order to check whether the componential hypothesis that is formulated with the item structure vectors can be found in the estimated parameters. Note that the (mean) values of the discrimination parameter are relative to the identification restrictions with respect to the variance of the person ability parameters and to the fixing of the maximum estimate at 1.5.

The *variance* of the discrimination parameters was rather high, the values of the standard deviations being about equal across tasks and dimensions (see Table 5). The high variability is probably due to the small sample size, resulting in the individual parameter estimates not being very

Table 5
Mean TESTMAP Discrimination Parameters (and Standard Deviations) for Model 2D_e.

Task	1	2
Generation	.31 (.31)	.38 (.35)
Evaluation	0	.33 (.32)
Open Synonym	.26 (.30)	.39 (.36)

reliable. Only the global solution should be interpreted and not the individual discrimination parameter estimates. In fact, the solutions obtained from TESTMAP (but also from NOHARM) even contained some negative values for the estimated discrimination parameters.

Correlation Between Generation and Evaluation

For the items of the generation subtask the weighted mean angle was calculated with the evaluation subtask dimension (cf. Equation 7) in order to get an estimate of the correlation between the generation subtask and the evaluation ability (cf. Equation 6). The weighted mean angle varied between 37° (adjectives of List 2) and 46° (verbs of List 1). Averaged over the six data sets, the weighted mean angle is 41° with a standard deviation of 3° for the means in the six data sets. This mean angle corresponds with a correlation of .75 between generation and evaluation. The estimate of the correlation coefficient between the two latent dimensions in model 2D_ge(r) by NOHARM was .87 on the average over the six data sets. Note that these correlations must be considered correlations between perfect measures (i.e. without measurement error), since they are defined from an angle or from the latent ability dimensions.

These correlations are too high for a reliable differentiation between generation and evaluation. The high correlation may be caused by a common underlying verbal ability. In fact, Janssen et al. (1996) found that in a structural equation solution for synonym tasks and reference tests, the evaluation factor, on which the evaluation and open synonym task loaded, correlated .57 with the generation-fluency factor, on which the generation task, the open synonym task, and two verbal fluency measures loaded. Another explanation of the high correlation can be found in the design of the present study: The same stimulus words were used for the three synonym tasks. In another study by Janssen and De Boeck (1996), the effect of common stimulus words was investigated using structural equation modeling. They found that repeating stimulus words across tasks enhanced the correlations among these tasks, especially when the tasks shared a common item format.

Item Difficulty Parameters

We also looked at the estimated item difficulty parameters in model 2D_e as obtained from TESTMAP. Again, the results were comparable with those obtained from NOHARM. On the average, the items of the evaluation subtask were the easiest, with a mean over the six data sets of -.66. The items of the generation subtask were most difficult (*M* = 1.00), closely followed by the items of the open synonym task (*M* = .89). The standard deviations of the item difficulty parameters of the items of one task varied between .72 (evaluation, adjectives of List 1) and 1.68 (open synonym, nouns of List 1).

Because the same stimulus words were used in the three tasks, the item difficulty of the open synonym task could be regressed on the item difficulty of the generation and evaluation subtask. The componential hypothesis implies that the open synonym task difficulty can be explained reasonably well from the generation and evaluation difficulties. The fact that the componential hypothesis can be tested also with respect to difficulties, is an advantage of componential IRT modeling in comparison with common factor-analytic methods. The resulting squared multiple correlation coefficients varied between .81 and .97 for the six data sets. The regression weight of the generation subtask item was always significant with $p < .001$. For the evaluation subtask items a significant regression weight was obtained for two of the six data sets with $p < .01$ and for four data sets with $p < .10$. These results indicate that in general both generation and evaluation contribute to the difficulty of the open synonym task. The fact that the contribution of evaluation is smaller can be understood from model 2D_e, as in that model generation also incorporates evaluation. Hence, the results of the regression analysis on the item difficulty are in line with the particular componential structure of the preferred model.

Discussion

As a general conclusion, one can say that the componential hypothesis of the synonym tasks was corroborated on the basis of the confirmatory M2P IRT model. The preferred model 2D_e showed a relatively good approximation to the data, with a mean γ_{ULS} statistic of .78 when averaged over the six data sets. According to that model, the generation subtask and the open synonym task turned out to be more similar in the abilities involved than was initially conceived of in the componential hypothesis. Also, the results of the item difficulty parameters were in line with this finding. Our results therefore indicate that the confirmatory M2P model can be considered a valuable tool in cognitive, componential research.

On a methodological level, several interesting results were obtained. First, a limitation of the study may be its small sample size. Indeed, the small sample size probably gave rise to less stable solutions at the item parameter level. However, given the cross validation with the six data sets and the correspondence between the two estimation methods, the global results seemed to be reliable. Hence, it seems that for research purposes, the confirmatory M2P model can be used even with a relatively small number of observations. Second, the present research confirms the robustness of the bivariate information methods in comparison with the full-information methods. Both estimation methods resulted in about equivalent parameter estimates. Moreover, the ordering of the confirmatory models by the goodness-of-fit measure of NOHARM was very consistent across data sets, while the AIC and CAIC showed more fluctuations across data sets, probably resulting from minor differences in the data sets. Third, it was shown in the present article that indicators of relative fit and absolute goodness of approximation can be combined for model selection. Finally, it was also shown how an estimate of a correlation could be obtained, even within an orthogonal solution.

The present article shows how multidimensional IRT models can complement both an experimental approach concentrated on the effect of task manipulations, and a factor-analytic approach aiming at modeling the structure in individual differences. Davison and Skay (1991) argued in favor of multidimensional scaling as a viable alternative to factor analysis for the analysis of person by item data in a multidimensional space. They indicated that multidimensional scaling stresses variation in task content, whereas factor analysis emphasizes (co)variation over individuals. The present article suggests that multidimensional IRT models can integrate both aspects. These models can be used to differentiate between different componential hypotheses in a multidimensional space, taking into account the structure of individual differences and without neglecting differences in item difficulty related to task content.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27-41). Amsterdam: North-Holland.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika, 52,* 345-370.

Butter, R., De Boeck, P., & Baele, J. (1992). *A combined cognitive componential and psychosemantic analysis of synonym tasks*. Unpublished manuscript, Department of Psychology, University of Leuven, Belgium.

Butter, R., De Boeck, P., & Verhelst, N. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika, 63,* 47-63.

Davison, M. L. & Skay, C. L. (1991). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin, 110,* 551-556.

Embretson (Whitely), S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3-17). New York: Academic Press.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-373.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fraser, C. (1988). *NOHARM: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, N.S.W., Australia: University of New England, Centre for Behavioral Studies.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

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

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

Janssen, R. & De Boeck, P. (1996). The contribution of a response-production component to a free-response synonym task. *Journal of Educational Measurement, 33,* 417-432.

Janssen, R. & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37-50.

Janssen, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence, 22,* 291-310.

Janssen, R., Hoskens, M., & De Boeck, P. (1993). An application of Embretson's multicomponent latent trait model to synonym tests. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology. Proceedings of the 7the European Meeting of the Psychometric Society in Trier* (pp. 187-190). Stuttgart: Gustav Fischer Verlag.

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149-176.

Knol, D. L. & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model: noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247-255.

McDonald, R. P. & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23-40.

McKinley, R. L. (1988, April). *Assessing dimensionality using confirmatory multidimensional IRT.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

McKinley, R. L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory* (Research Report No. 89-21). Princeton, NJ: Educational Testing Service.

McKinley, R. L. (1992). *TESTMAP Version 2.1 User's Guide*. Unpublished manuscript.

McKinley, R. L. & Kingston, N. M. (1988, April). *Confirmatory analysis of test structure using multidimensional IRT*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

McKinley, R. L. & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389-390.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Muthén, B. (1984). Contributions to factor analysis of dichotomized variables. *Psychometrika, 43*, 551-560.

Reckase, M. D. & McKinley, R. L. (1982). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 151-177). Unpublished manuscript, Minneapolis, University of Minnesota, Department of Psychology.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10 - 39). Newbury Park, CA: Sage.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika, 48*, 259-267.

van Leeuwe, J. F. J. & Roskam, E. E. (1991). The conjunctive item response model: A probabilistic extension of the Coombs and Kao model. *Methodika, 5*, 14-32.

Wilson, D. T., Wood, R., & Gibbons, R. T. (1984). *TESTFACT. Test scoring, item statistics, and item factor analysis*. Mooresville, IN: Scientific Software.

Accepted August, 1998.