Multidimensional Componential Item Response Theory Models for Polytomous Items

Machteld Hoskens and Paul De Boeck

Applied Psychological Measurement 2001 25: 19

DOI: 10.1177/01466216010251002

The online version of this article can be found at:

<http://apm.sagepub.com/content/25/1/19>

Published by:

<http://www.sagepublications.com>

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: <http://apm.sagepub.com/cgi/alerts>

Subscriptions: <http://apm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://apm.sagepub.com/content/25/1/19.refs.html>

>> [Version of Record](#) - Mar 1, 2001

[What is This?](#)

Multidimensional Componential Item Response Theory Models for Polytomous Items

Machteld Hoskens and Paul De Boeck
University of Leuven, Belgium

A framework for modeling componential data using item response theory models for polytomous items is presented. This framework models response accuracies on complex cognitive tasks, which are decomposed in terms of more basic elements (components), such as knowledge structures, cognitive processes, and strategies. Response categories of polytomous items correspond to possible response patterns on the components. Models with only single component effects are distinct from models that allow for component combination effects. For the

latter, a further distinction is made between constant and dimension-dependent component combination effects as a function of whether these effects depend on an examinee's ability. The proposed models are multidimensional, because separate abilities are assumed to be involved in their various components. An example illustrates the approach. *Index terms: cognitive research, componential models, item response theory, local item dependence, multidimensionality, synonym tasks, test design, testlets.*

Several models have been proposed that integrate cognitive theorizing with the item response theory (IRT) framework (e.g., Embretson, 1980, 1984, 1995; Fischer & Formann, 1982; Jannarone, 1986; Kelderman & Rijkes, 1994; Maris, 1995; Mislevy & Verhelst, 1990; Mislevy & Wilson, 1996; Wilson, 1989; Yamamoto & Gitomer, 1993). IRT can be used as a tool to investigate the cognitive nature of complex tasks and the extent to which various knowledge structures, cognitive processes, and strategies are involved in solving those tasks. This type of decomposition of a complex cognitive task in terms of more basic elements can be referred to as *componential research* (Sternberg, 1985, 1997). (Used in a broader sense here, the term "component" denotes all basic elements considered, rather than restricting it to cognitive processes only.)

The componential nature of a task is modeled here by considering all possible response patterns on the components postulated for the task. They are then modeled as the response categories of a polytomous item. For example, there are four possible response patterns for two components: (1) no correct components, (2) the first component correct, (3) the second component correct, or (4) both components correct. Therefore, the IRT model is for a polytomous item with four response categories. Modeling componential data as polytomous items is advantageous, because single and combination component effects can be taken into account.

Hoskens & De Boeck (1995) presented this approach in a *unidimensional* framework. Here, their framework is extended to the *multidimensional* case—solving each component involves a separate ability. A similar approach was taken by Fischer (1973a) for modeling speed and accuracy simultaneously. Speed and accuracy were combined into three categories: (1) a correct and fast response, (2) a correct and slow response, and (3) an incorrect response. Two ability dimensions were used to model the three categories. There are two types of componential data that can be

modeled in the framework for polytomous items: componential testlet data and genuine polytomous item data (Hoskens & De Boeck, 1995). Both types are described below.

In componential research, the subtask paradigm can be used to develop subtasks, in order to isolate components that are postulated to underlie a complex task. Componential subtasks belonging to the same item stem can then be taken together as a testlet (Wainer & Kiely, 1987). This is called a *componential testlet*. The term “testlet” [also “subtest” (Andrich, 1985) or “item bundle” (Rosenbaum, 1988)] refers to a case in which an item set is considered a unit to be modeled, rather than just a single item. In a testlet, response patterns for the item set are modeled by treating them as categories of a polytomous item.

The testlet approach can also be used in other contexts than componential research; for example, it can be used to model different reading comprehension items that relate to the same text. The advantage of the testlet approach is that local dependence between the items constituting the testlet can be taken into account (i.e., local dependence between subtasks). A number of researchers have addressed local item dependence (e.g., Andrich, 1985; Rosenbaum, 1988; Steinberg & Thissen, 1996; Thissen & Steinberg, 1988; Thissen, Steinberg, & Mooney, 1989; Wilson, 1988; Wilson & Adams, 1995; Yen, 1984, 1993). Jannarone’s (1986) conjunctive kernel model can be viewed as a testlet model—the conjunctions of items between which local dependence is modeled can be viewed as testlets. The present paper describes an example of this kind.

A second context in which componential data can be analyzed using models for polytomous items involves starting with a task analysis and developing real polytomous items with componentially designed response alternatives. Each alternative is assumed to reflect a different success or failure pattern on the components. Although the models described below apply to this type of data, it is not further examined here.

The approach presented here offers a formal framework that can easily be applied beyond the context of componential research to any context in which local dependence occurs between items, such as in achievement tests and personality or attitude questionnaires (e.g., Hoskens & De Boeck, 1997). Through local item dependence, interesting phenomena of the data can be revealed (e.g., learning effects and response consistency effects).

The testlet approach is not the only approach that has been used to address local item dependence. For example, virtual (technical) items have also been used (Verhelst & Glas, 1993). A virtual item is an item that exists only in a formal sense and not as a separate item in the test. Virtual items usually model subsets of responses given under certain conditions to a real item. For example, a real item on a test can be modeled as two virtual items: one virtual item for responses given following a correct response on the previous item, and one for responses given following an incorrect response. Other approaches suggested for addressing local item dependence are hierarchical IRT models (Patz, 1996) and Markov chain models (Ackerman & Spray, 1986).

A hypothesis testing framework is presented for investigating the presence and nature of local item dependence. This framework complements other, more exploratory approaches (e.g., Chen & Thissen, 1997; Yen, 1984, 1993). Local item dependencies are combination effects that are symmetrical in nature. However, the approach can be generalized for modeling component combinations (or more generally, item combinations, when the framework is applied outside of the context of componential research) that are affected by item order (Hoskens & De Boeck, 1997).

Psychometric Models

In Bock’s (1972) nominal model, the probability of an examinee with a trait level (θ) selecting response alternative m on a polytomous item is

$$P(m) = \frac{\exp[z_m(\theta)]}{\sum_{n=1}^M \exp[z_n(\theta)]}, \quad (1)$$

with $n = 1, 2, \dots, m, \dots, M$, and response function

$$z_n(\theta) = a_n\theta + c_n, \quad (2)$$

where a_n is the category discrimination parameter (for the capability of discriminating between examinees with different θ s) and c_n is the category threshold.

To obtain model identifiability, reparameterize as follows:

$$a' = \alpha' \mathbf{T}_a, \quad (3)$$

and

$$c' = \gamma' \mathbf{T}_c, \quad (4)$$

where α and γ are vectors of length $(M - 1)$ containing unconstrained parameters, and $\mathbf{T}_a$ and $\mathbf{T}_c$ are transformation matrices of order $(M - 1) \times M$ that link the unconstrained parameters to the response categories of the polytomous item. Through $\mathbf{T}_a$ and $\mathbf{T}_c$, the models presented below are written as variants of Bock's (1972) nominal model (cf. Thissen & Steinberg, 1986). These transformation matrices indicate the componential design of the polytomous item.

Examinee responses to two components (denoted C1 and C2) are considered. Both components reflect a separate θ . In this respect, the approach presented here differs from that described by Hoskens & De Boeck (1995), where components were assumed to depend on the same underlying θ . Considering all possible response patterns on two dichotomously scored components, a polytomous item is obtained with four response categories: (1) no components correct, (2) only C1 correct, (3) only C2 correct, and (4) both components correct. The polytomous item can either correspond to a real polytomous item with componentially designed response alternatives or to a testlet resulting from combining subtask data. This can easily be generalized to polytomous items reflecting examinee responses to more than two components.

Single Component Effects Model

In the single component effects model (Model 1), the probability of the response pattern corresponding to a correct response on both components is the product of the probabilities for answering each of the separate components correctly. Therefore, there is local independence between the components. A separate Rasch scale is posited for each of the components. Writing the probabilities for each of the four possible response patterns yields

C1	C2	Probability	
0	0	$1/\nu(\theta)$	
1	0	$\exp(\alpha_1\theta_1 - \gamma_1)/\nu(\theta)$	(5)
0	1	$\exp(\alpha_2\theta_2 - \gamma_2)/\nu(\theta)$	
1	1	$\exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2)]/\nu(\theta)$,	

with

$$\nu(\theta) = 1 + \exp(\alpha_1\theta_1 - \gamma_1) + \exp(\alpha_2\theta_2 - \gamma_2) + \exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2)], \quad (6)$$

where γ_1 and γ_2 are item-specific parameters that indicate the difficulty of the first and the second components, respectively, for this testlet item.

The first two columns of Equation 5 reflect the response patterns on the components, with 1 indicating a correct response and 0 indicating an incorrect response. The third column gives the probabilities for each of the four response patterns. The difficulty of having both components correct equals the sum of γ_1 and γ_2 .

For each θ dimension, α and $\mathbf{T}_a$ are

$$\alpha_1 \text{ and } \mathbf{T}_a = \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix}, \quad (7)$$

and

$$\alpha_2 \text{ and } \mathbf{T}_a = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}. \quad (8)$$

The person θ weights α_1 and α_2 are constant over items. The matrix $\mathbf{T}_a$ indicates the extent to which the weighted person θ s are involved in each of the possible response patterns. Only one component θ is involved in response patterns corresponding to a single correct response to only one of the components, whereas both component θ s are involved in the response pattern corresponding to having both components correct.

For the category thresholds,

$$\boldsymbol{\gamma}' = [\gamma_1 \quad \gamma_2], \quad (9)$$

and

$$\mathbf{T}_c = \begin{bmatrix} 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & -1 \end{bmatrix}. \quad (10)$$

$\mathbf{T}_c$, when multiplied by -1 , equals the transpose of the matrix formed by the first two columns of Equation 5, which indicates the componential design of the polytomous item. (This is also the case for the models presented below.)

Constant Component Combination Effects (CCE) Model

In this model with constant CCE (Model 2), there is dependency between components. The difficulty of having both components correct is increased or decreased in comparison with Model 1. It is assumed that the increase or decrease is constant over examinees differing in θ . There is local dependence between the components. Considering examinee responses to two components, the following probabilities result:

C1	C2	C1 $\times$ C2	Probability	
0	0	0	$1/\nu(\theta)$	
1	0	0	$\exp(\alpha_1\theta_1 - \gamma_1)/\nu(\theta)$	(11)
0	1	0	$\exp(\alpha_2\theta_2 - \gamma_2)/\nu(\theta)$	
1	1	1	$\exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2) - \gamma_{12}]/\nu(\theta)$,	

with

$$\nu(\theta) = 1 + \exp(\alpha_1\theta_1 - \gamma_1) + \exp(\alpha_2\theta_2 - \gamma_2) + \exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2) - \gamma_{12}]. \quad (12)$$

The CCE is denoted in the third column of Equation 11. This column can be obtained by taking the product of the columns for the single component effects, Columns 1 and 2 (note that the probability

of two correct answers does not equal the product of the probabilities of the two separate correct answers).

In comparison to Model 1, an extra parameter is added to the numerator of the response pattern corresponding to both components correct. This accounts for the CCE. Consequently, the denominator of the polytomous item is also changed. Therefore, the parameter for the CCE affects all response patterns of the testlet, not only the response pattern reflecting a joint correct response. The extra parameter, γ_{12} , indicates the extent to which a joint correct response to both components is easier or more difficult than expected taking only the effects of single components into account. Given that all other parameters are the same as those of Model 1, $\gamma_{12} < 0$ indicates that a joint correct response to both components is easier than predicted on the basis of the separate difficulties of the components, and $\gamma_{12} > 0$ indicates that it is more difficult.

Correspondingly,

$$\boldsymbol{\gamma}' = [\gamma_1 \quad \gamma_2 \quad \gamma_{12}] , \quad (13)$$

and

$$\mathbf{T}_c = \begin{bmatrix} 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix} . \quad (14)$$

The third row of $\mathbf{T}_c$ indicates that γ_{12} is added to the numerator when writing the probability for the response pattern of answering both components correctly (and also to the denominator for all response patterns). α_1 and α_2 and their corresponding transformation matrices are the same as for Model 1.

CCE Depending on θ Model

In this model with CCE depending on θ (Model 3), the CCE is no longer constant over examinees differing in θ , but depends on the θ levels of an examinee. There is local dependence between components. The probabilities for each of the four response patterns are

C1	C2	C1 $\times$ C2	Probability	
0	0	0	$1/\nu(\boldsymbol{\theta})$	
1	0	0	$\exp(\alpha_1\theta_1 - \gamma_1)/\nu(\boldsymbol{\theta})$	(15)
0	1	0	$\exp(\alpha_2\theta_2 - \gamma_2)/\nu(\boldsymbol{\theta})$	
1	1	1	$\exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2) + (\alpha_1\theta_1 + \alpha_2\theta_2 - \gamma_{12})]/\nu(\boldsymbol{\theta})$,	

with

$$\begin{aligned} \nu(\boldsymbol{\theta}) = & 1 + \exp(\alpha_1\theta_1 - \gamma_1) + \exp(\alpha_2\theta_2 - \gamma_2) + \exp[(\alpha_1\theta_1 - \gamma_1) + (\alpha_2\theta_2 - \gamma_2) \\ & + (\alpha_1\theta_1 + \alpha_2\theta_2 - \gamma_{12})] . \end{aligned} \quad (16)$$

In addition to γ_{12} , the weighted person θ s are also added to the model to account for the CCE. Therefore, whether answering both components correctly is easier or more difficult than expected based on the difficulties of the separate components depends on an examinee's component θ s. Suppose that apart from γ_{12} , all parameters are the same as those of Model 1. Then, for examinees with $(\alpha_1\theta_1 + \alpha_2\theta_2) > \gamma_{12}$, answering both components correctly is easier than expected. The reverse is the case for examinees with $(\alpha_1\theta_1 + \alpha_2\theta_2) < \gamma_{12}$. Model 3 can be seen as a multidimensional generalization of Jannarone's (1986) locally dependent conjunctive measurement model.

Correspondingly,

$$\alpha_1 \text{ and } \mathbf{T}_a = \begin{bmatrix} 0 & 1 & 0 & 2 \end{bmatrix}, \quad (17)$$

and

$$\alpha_2 \text{ and } \mathbf{T}_a = \begin{bmatrix} 0 & 0 & 1 & 2 \end{bmatrix}. \quad (18)$$

γ and $\mathbf{T}_c$ are as given in Equations 13 and 14 for Model 2.

Graphical Display of the Models

The effect of the CCE and, particularly, the difference between Models 2 and 3, can be seen in a graphical presentation of the probability of each of the response patterns as a function of θ . Because two θ s are involved, three-dimensional plots are obtained. Figures 1, 2, and 3 show the response surfaces of a hypothetical example for each of the four possible response patterns for Models 1, 2, and 3, respectively (where 1 = a correct response and 0 = an incorrect response). The item parameter values used were 0, -1 , and -1.5 for γ_1 , γ_2 , and γ_{12} , respectively. The person θ weights for both dimensions, α_1 and α_2 , were set equal to 1. For all three models and each of the response patterns, the response surfaces reflect nonzero probabilities mainly in one quadrant of the plane formed by the two θ dimensions.

For Model 2 (Figure 2), the response surfaces are shifted compared to the response surfaces for Model 1 (Figure 1). For the response pattern with both components correct (1,1; Figure 2d), the response surface is shifted toward the lower left of the horizontal plane compared to its counterpart in Figure 1d. Hence, compared to Model 1, for lower levels of either of the two θ s, a high probability is obtained for a 1,1 response pattern when there is a constant CCE.

For Model 3 (Figure 3), the steepness of the response surfaces is different from that of Model 1. For example, for the 1,1 response pattern (Figure 3d), cross-sections of this surface for values of either θ would show steeper functions for Model 3 than for Model 1 (Figure 1d). This increased steepness is caused by adding the weighted person θ s to the response function of the response pattern to account for the CCE.

Model 3 is only one variant of a larger family of models with combination effects depending on θ . For example, the weighted person θ s can be added two or more times to account for the CCE, increasing the steepness of the response surfaces even further. The weighted θ s could also be subtracted, yielding a decrease in steepness. This would mean that the CCE is inversely related to one or both of the dimensions accounting for examinee responses to single items. Because a combination effect in the context of educational assessment might indicate learning effects, this would then indicate that learning is greater for low-ability examinees than for high-ability examinees.

In Figure 3, there is a trade-off between the two θ s in the combination effect—a high θ in solving one component can compensate for a low θ in solving the other component [such that the weighted sum of both is greater (or smaller) than γ_{12}]. Another multidimensional model with CCE depending on θ would have the CCE depend only on one θ . For example, suppose the CCE depended only on the ability to solve the first component. The term added to account for the CCE when writing the probability of having both components correct would be $(\alpha_1\theta_1 - \gamma_{12})$. Then, given that all parameters except γ_{12} are equal to those of Model 1, having both components correct would be easier than expected for examinees with $\alpha_1\theta_1 > \gamma_{12}$. The reverse would be true for examinees with $\alpha_1\theta_1 < \gamma_{12}$. The magnitude and direction of the CCE now would vary as a function of an examinee's position on Dimension 1 only.

Figure 2
Response Surfaces for Two-Dimensional Model 2

Figure 3
 Response Surfaces for Two-Dimensional Model 3

A further generalization of Model 3 would be a model that estimates the extent to which the CCE depends on θ . A multidimensional variant of the two-parameter logistic model (Birnbau, 1968) would be obtained. When examinee responses to two components are considered, there could be, for example, two discrimination parameters estimated for each component—one indicating the extent to which a single correct response to one component depends on the component θ_1 , and a second indicating the extent to which the CCE depends on θ_2 .

Estimation of Model Parameters

Two types of programs can be used to estimate parameters of these multidimensional models. The quasi-loglinear approach for estimating IRT models can be implemented using LOGIMO (Kelderman & Steen, 1993) or any general computer program for loglinear analysis. Conditional maximum likelihood estimates of the item parameters can then be obtained. In LOGIMO, item parameter estimates are obtained iteratively using the Deming-Stephan algorithm (Bishop, Fienberg, & Holland, 1975; Deming & Stephan, 1940) and the Newton-Raphson algorithm (for more details and for maximizing the likelihood functions, see Kelderman, 1996). Model 1 then can be estimated by modeling the components as separate items and relating the various types of components to separate dimensions.

Model 2 also can be estimated by considering the components as separate items. The CCE is modeled by adding item interaction parameters to the model. In addition, Model 2 can be estimated by considering polytomous items. The response alternatives of the polytomous item reflect the possible response patterns on components that belong to the same item stem. By means of reparameterization, parameter estimates are obtained for single components and for CCE. When using LOGIMO, scoring weights that indicate the extent to which each θ is involved in the response categories need to be specified. These scoring weights correspond to T_a .

Model 3 differs from Model 2 only in terms of these transformation matrices for category discrimination. When using LOGIMO, Model 3 is therefore also estimated by considering polytomous items (with specification of the correct scoring weights) and by using reparameterization to obtain estimates of effects for single components and for component combinations. When using a general program for loglinear analysis, all three models can be estimated by considering the components as separate items.

CONQUEST (Wu, Adams, & Wilson, 1997; see also Adams, Wilson, & Wang, 1997) also can be used to estimate model parameters. CONQUEST allows for estimation of various unidimensional and multidimensional models for polytomous items with a priori specified weights for the θ s. For each response category, the relevant θ can be specified. A design matrix is used for specifying the composition of the category thresholds in terms of more basic parameters. CONQUEST assumes that θ is random and has a multivariate normal distribution over dimensions, with estimated variances and covariances. As a result, marginal maximum likelihood estimates of the item parameters are obtained. Item parameter estimates can be obtained with the Bock & Aitkin (1981) formulation of the EM algorithm, using quadrature with a fixed grid of nodes, or using monte carlo sampling. (For more details and for the likelihood functions to be maximized, see Adams et al., 1997; Wu et al., 1997.)

Using marginal maximum likelihood, CONQUEST avoids a common problem found using LOGIMO or a general program for loglinear analysis. In LOGIMO, score group parameters need to be estimated in order to achieve the conditioning on the sufficient statistics for θ . The number of these parameters rapidly increases as the numbers of items and dimensions increase, or as the extent to which θ is involved in the various response categories of the items increases. For example, when five polytomous testlet items are modeled using a two-dimensional Model 2, in principle, 36 score

groups can occur (the score for each component can range between 0 and 5, and every combination of component scores is possible). When five polytomous testlet items are modeled using a two-dimensional Model 3, 52 score groups can occur in principle.

The availability of estimation algorithms does not guarantee that the models to be estimated are identifiable given the data to which they are applied. Even for the unidimensional linear logistic test model (Fischer, 1973b), the identifiability conditions are rather complicated (Fischer, 1995). The multidimensional componential IRT models proposed here are a class of models that can be estimated using LOGIMO and CONQUEST. However, the identifiability of the componential models and of the more general multidimensional modeling frameworks at the basis of these two programs deserves further attention.

Assessment of Goodness of Fit

To assess whether accounting for CCE increases the goodness of fit of a model to the data, two measures of relative goodness of fit can be used. The likelihood-ratio-test statistic can be used when models are hierarchically related, as are Model 1 and Model 2. Akaike's (1977) information criterion (AIC) can be used when models are not hierarchically related, as is the case when comparing Model 3 to either Model 1 or 2, or when comparing a variation of Model 3 differing in the degree to which the CCE depend on each of the θ s. AIC can also be used to assess whether positing a separate dimension for each component results in improved goodness of fit of a model to the data, rather than relating all components to the same underlying θ .

Example

Method

To demonstrate the approach, an example was taken from a study in which the componential nature of examinee responses to synonym tasks was investigated (Janssen, 1994; Janssen & De Boeck, 1997). Two components were considered as underlying examinee responses to a synonym task with open-ended questions (e.g., "Give a synonym for..."): (1) a generation component (e.g., generating synonyms for a given stimulus word), and (2) an evaluation component (e.g., evaluating synonyms with respect to their correctness as a synonym for the given stimulus word). In an attempt to isolate the two components, the subtask paradigm was used. Two subtasks were formulated, corresponding to the generation component and the evaluation component, respectively. In the *generation subtask*, examinees were asked to write down all words that came to mind when searching for a synonym of a given stimulus word. Examinee responses were scored correct if the correct synonym was among the words generated. In the *evaluation subtask*, examinees indicated in a list of four response alternatives those words they thought to be true synonyms of a given stimulus word. The alternatives for each item were: the correct synonym, determined by a general dictionary of the Dutch language (Geerts & Heestermans, 1984), and the words most frequently given in the general subtask for the given stimulus word. Examinee responses on the evaluation subtask were scored as correct if the correct synonym was indicated. There was a time interval of at least six weeks between the two tasks, in order to avoid memory effects.

The stimulus material used in both subtasks consisted of an item set divided into two lists. Each list contained three subsets: verbs, adjectives, and nouns. The two lists were administered to 212 and 257 examinees, respectively, all fifth and sixth grade humanities students. Each of these six datasets was modeled separately to determine if the results replicated. Only items with nonzero frequencies for each of the four possible response patterns on the subtasks were modeled. For each of the six datasets, there were between 11 and 19 items (see Table 1).

Table 1
Number of Items (k), and Mean (M) and Variance (Var) of Percent Correct for Synonym Task Data

Item Set	Generation			Evaluation		
	k	M	Var	M	Var	N
List 1 ($N = 212$)						
Verbs	18	.28	.06	.66	.05	212
Adjectives	15	.27	.05	.69	.06	212
Nouns	11	.25	.06	.61	.08	212
List 2 ($N = 257$)						
Verbs	18	.41	.07	.68	.04	257
Adjectives	19	.34	.05	.66	.04	257
Nouns	16	.32	.05	.70	.08	257

The data were examined to determine the extent to which there was dependency between the two subtasks; that is, the extent to which a correct response to both subtasks that shared the same item stem (the same stimulus word) was easier or more difficult than would be predicted on the basis of the separate difficulties of the subtasks. Also investigated was the extent to which generation and evaluation in the two subtasks could be considered as separate abilities rather than as reflections of the same general verbal ability. Examinee responses to subtasks that shared the same item stem were considered a testlet and modeled as a polytomous item based on the four response patterns.

Results

Descriptive statistics for the two subtasks are given in Table 1. Table 2 shows results for the two-dimensional models: a model depending on both generation and evaluation (Model G/E), a model with CCE depending on generation only (Model G), and a model with CCE depending on evaluation only (Model E). CONQUEST (Wu et al., 1997) was used to estimate the models. In addition to the item means and variances of the item parameter estimates, Table 2 also shows estimates for the variances of E and G (Var_E and Var_G) and the covariance between G and E (Cov_{GE}). Model identification was obtained by constraining the mean θ to zero.

For all six datasets, a significant improvement in goodness of fit of the model to the data was obtained when both single component effects (constant) and CCE were included in the model [likelihood-ratio $\chi^2(18) = 110.6$, $\chi^2(15) = 146.4$, $\chi^2(11) = 111.6$ for verbs, adjectives, and nouns, respectively, using List 1; and $\chi^2(18) = 185.4$, $\chi^2(19) = 179.6$, $\chi^2(16) = 133.7$ for verbs, adjectives, and nouns, respectively, using List 2 (all $p < .01$)].

The AIC values indicate that allowing the CCE to be dimension dependent (on either C, E, or both) did not result in improved goodness of fit. Only for nouns using List 2 was the AIC value for Model E (372.7) smaller than for the constant model (374.3), but the difference was small. On average, for all six datasets, a correct response to both componential subtasks was easier than predicted from the single component model. The average combination effect parameter, γ_{GE} , added to account for the CCE, varied between $-.66$ for verbs and -1.39 for adjectives using List 1. For all six datasets, the correlation between the two θ dimensions decreased when constant CCE were taken into account [mean correlations were .48 for Model 1 (None) and .15 for Model 2 (Constant)].

Table 3 shows results obtained when the data were modeled unidimensionally using CONQUEST (Wu et al., 1997). As was the case for the two-dimensional models, for all six datasets, goodness of fit significantly improved when constant CCE were modeled for all six cases. Likelihood-ratio statistics were $\chi^2(18) = 93.6$, $\chi^2(15) = 135.4$, and $\chi^2(11) = 102.8$ for List 1 verbs, adjectives, and

nouns, respectively; and $\chi^2(18) = 155.1$, $\chi^2(19) = 151.6$, and $\chi^2(16) = 106.0$ for List 2 verbs, adjectives, and nouns, respectively (all $p < .01$).

For three of the six datasets, no further improvement in goodness of fit was obtained when the CCE were allowed to be dimension dependent. In terms of AIC, the best-fitting model appeared to be the model with constant CCE. For two datasets, the best-fitting model was the dimension-dependent (DD) model (however, the difference for the model with constant CCE was small). For List 2 nouns, the best-fitting model was the DD model. Overall, when modeling the subtask data unidimensionally, models with CCE constant over examinees differing in θ seemed to suffice. γ_{GE} varied between $-.50$ for verbs and -1.30 for adjectives, both using List 1. A joint correct response for both G and E was easier than expected based on the difficulties of the separate components.

Goodness of fit for the unidimensional models can be compared to that of the two-dimensional models. Almost all two-dimensional models had smaller AIC values than their unidimensional counterparts (exceptions were the model with no CCE and the DD model for List 1 adjectives). The better fit of the two-dimensional models shows that distinguishing between the two components (G and E) as separate traits was useful.

An indication of absolute goodness of fit of the two-dimensional model with constant CCE to the data was obtained by considering the approximate standard-normal-fit statistics provided by CONQUEST for each estimated γ , based on standardized weighted residuals (Wu, 1997). No dataset in List 1 had extreme positive values greater than 2.0. In List 2, values greater than 2.0 occurred only once for adjectives and nouns, and were not observed for verbs. t statistics with extreme negative values (< -2) were observed only twice for List 2 adjectives, twice for List 1 nouns, and once for List 1 adjectives. They were not observed for any other dataset. The fit statistics therefore did not suggest that the two-dimensional model with constant CCE should be rejected—the majority of the fit statistics fell well within the $[2, 2]$ range. These results suggest that there was dependence between both operations—generating a synonym and evaluating synonyms as to their correctness—when applied to the same item stem. There was no evidence for CCE that depended on the abilities to generate and evaluate candidate synonyms.

Discussion and Conclusions

Combination effects could indicate an overlap between the generation and evaluation of candidate synonyms, a result that was also found by Janssen (1994; Janssen & De Boeck, 1997). Combination effects found for the two components of the synonym task appeared to be constant over examinees who differed in ability. This result can be compared to a result found by Hoskens & De Boeck (1995) for a spelling task in Dutch. Two components were considered in the spelling of open and closed syllables: (1) writing the correct number of vowels, and (2) writing the correct number of consonants. Hoskens and De Boeck found CCE that depended on ability. A possible explanation for the difference between their study and the present study might be the difference in distribution of the various abilities over examinees. Because the examinees for the synonym task were all humanities students, it could be assumed that they had, at minimum, a moderate level of generation and evaluation ability. The spelling task measured what the examinees were learning in class, and a distinction could be made between those not yet mastering the rule for spelling open and closed syllables and those that could correctly apply this rule. Ability-dependent CCE then might also be found for the synonym task if there were more variation as to the examinees' generation and evaluation abilities.

The modeling framework presented here can be generalized in multiple ways. In addition to first-order combination effects between components, higher-order combination effects could also be taken into account. Each of the components could be related to a separate θ dimension, or some of

Table 2
Parameter Estimates and Goodness-of-Fit Statistics for Two-Dimensional Models
With No CCEs (None, or Model 1), Constant CCEs (Model 2), CCEs Dependent
on Both G and E (Model G/E), G Only (Model G), and E Only (Model E)

Model	Items						Persons			Npar	-2logL	AIC*
	γ_G		γ_E		γ_{GE}		Var_G	Var_E	Cov_{GE}			
	M	Var	M	Var	M	Var						
List 1												
Verbs												
None	1.41	2.75	-1.00	2.12			.38	.35	.12	39	7312.9	390.9
Constant	1.85	3.13	-.88	2.39	-.66	.81	.37	.33	.01	57	7202.3	316.3
Model G	1.80	3.02	-.87	2.38	-.55	.79	.13	.32	-.01	57	7221.7	335.7
Model E	2.01	3.10	-.86	2.34	-.81	.80	.52	.19	-.16	57	7212.2	326.2
Adjectives												
None	1.47	2.05	-1.06	2.09			.69	.36	.36	33	5894.6	460.6
Constant	2.42	3.04	-.78	2.19	-1.39	1.78	.58	.20	.20	48	5748.2	344.2
Model G/E	2.44	2.75	-.75	2.15	-1.40	1.70	.15	.15	-.02	48	5759.4	355.4
Model G	2.37	2.80	-.77	2.18	-1.37	1.73	.15	.19	.06	48	5753.9	349.9
Model E	2.48	2.89	-.77	2.16	-1.43	1.72	.40	.15	.05	48	5755.4	351.4
Nouns												
None	1.79	3.35	-.67	2.77			.57	.25	.20	25	4207.6	257.6
Constant	2.62	3.23	-.45	2.61	-1.24	1.60	.52	.16	.05	36	4096.0	168.0
Model G/E	2.60	2.97	-.44	2.58	-1.22	1.59	.18	.15	-.08	36	4101.4	173.4
Model G	2.53	3.03	-.45	2.59	-1.13	1.60	.14	.16	.02	36	4100.8	172.8
Model E	2.74	3.12	-.45	2.60	-1.36	1.59	.58	.14	-.09	36	4097.3	169.3

continued on next page

Table 2, continued
Parameter Estimates and Goodness-of-Fit Statistics for Two-Dimensional Models
With No CCEs (None, or Model 1), Constant CCEs (Model 2), CCEs Dependent
on Both G and E (Model G/E), G Only (Model G), and E Only (Model E)

Model	Items						Persons			Npar	−2logL	AIC*
	γ_G		γ_E		γ_{GE}		Var_G	Var_E	Cov_{GE}			
	M	Var	M	Var	M	Var						
List 2												
Verbs												
None	.81	3.73	−.96	1.21			.33	.40	.15	39	9531.0	609.0
Constant	1.56	4.27	−.63	1.16	−1.08	.52	.29	.38	.00	57	9345.6	459.6
Model G	1.52	4.17	−.61	1.14	−1.05	.50	.12	.40	−.04	57	9374.5	488.5
Model E	1.71	4.22	−.60	1.10	−1.22	.51	.42	.18	−.15	57	9366.0	480.0
Adjectives												
None	1.00	2.09	−.81	1.30			.37	.41	.17	41	10239.1	321.1
Constant	1.72	2.52	−.54	1.23	−1.09	.67	.33	.39	.03	60	10059.5	179.5
Model G/E	1.81	2.36	−.51	1.18	−1.18	.63	.17	.30	−.15	60	10075.6	195.6
Model G	1.68	2.43	−.52	1.21	−1.03	.64	.12	.38	−.01	60	10088.5	208.5
Model E	1.90	2.45	−.51	1.16	−1.26	.64	.46	.22	−.16	60	10064.8	184.8
Nouns												
None	1.17	2.04	−1.38	3.55			.62	.34	.16	35	7606.0	476.0
Constant	2.07	2.39	−1.09	3.33	−1.28	.80	.63	.30	−.02	51	7472.3	374.3
Model G/E	2.08	2.15	−1.07	3.29	−1.28	.74	.27	.27	−.16	51	7477.1	379.1
Model G	1.95	2.19	−1.08	3.31	−1.15	.76	.17	.30	−.03	51	7479.3	381.3
Model E	2.27	2.28	−1.07	3.27	−1.48	.76	.82	.21	−.20	51	7470.7	372.7

Note. No results were available due to convergence failures for Verbs Lists 1 and 2 for Model G/E.

*A constant was subtracted for models applied to the same data.

Table 3
Parameter Estimates and Goodness-of-Fit Statistics for Unidimensional Models
With No CCEs (None), Constant CCEs, and a Dimension-Dependent Model (Model DD)

Model	Items						Persons Var	Npar	−2logL	AIC*
	γ_G		γ_E		γ_{GE}					
	M	Var	M	Var	M	Var				
List 1										
Verbs										
None	1.38	2.65	−.98	2.05			.24	37	7337.5	411.5
Constant	1.71	3.06	−.90	2.35	−.50	.81	.19	55	7243.9	353.9
Model DD	1.73	3.06	−.87	2.33	−.40	.78	.12	55	7276.7	386.7
Adjectives										
None	1.41	1.90	−1.07	2.12			.41	31	5899.8	461.8
Constant	2.30	2.91	−.82	2.25	−1.30	1.78	.28	46	5764.4	356.4
Model DD	2.32	2.82	−.80	2.15	−1.30	1.71	.14	46	5762.9	354.9
Nouns										
None	1.72	3.12	−.68	2.80			.27	23	4212.3	258.3
Constant	2.47	3.05	−.48	2.67	−1.13	1.60	.16	34	4109.5	177.5
Model DD	2.47	3.06	−.48	2.66	−1.03	1.58	.12	34	4108.9	176.9
List 2										
Verbs										
None	.80	3.64	−.94	1.15			.24	37	9558.5	632.5
Constant	1.44	4.22	−.66	1.14	−.91	.52	.17	55	9403.4	513.4
Model DD	1.44	4.21	−.66	1.13	−.86	.49	.11	55	9467.5	577.5
Adjectives										
None	.99	2.03	−.79	1.25			.27	39	10271.4	349.4
Constant	1.60	2.50	−.56	1.19	−.93	.66	.20	58	10119.8	235.8
Model DD	1.61	2.47	−.55	1.16	−.88	.63	.12	58	10159.6	275.6
Nouns										
None	1.13	1.87	−1.36	3.51			.30	33	7650.6	516.6
Constant	1.83	2.30	−1.14	3.40	−1.04	.78	.22	49	7544.6	442.6
Model DD	1.87	2.24	−1.11	3.40	−1.00	.74	.12	49	7525.8	423.8

*A constant was subtracted for models applied to the same data.

the components could be considered as reflecting the same underlying θ s. In the multidimensional models presented here, the extent to which CCE depended on θ was a priori specified. In an analogous manner, such as when modeling data unidimensionally, there could be models in which the extent to which CCE are θ dependent is estimated. Effects for combinations of components that are influenced by order effects could also be estimated.

Several types of componential models can be formulated. Two such models are:

1. A model based on the decomposition of either the item parameter, the person parameter, or both. The linear logistic test model (Fischer, 1973b) is an example of a model in which the item difficulty parameter is decomposed in terms of contributions of more basic components. In multidimensional compensatory IRT models (Adams et al., 1997; Kelderman & Rijkes, 1994), the person parameter is decomposed into a weighted sum of dimension-specific person parameters, and the set of θ dimensions required to solve the item can vary between items. Spada & McGaw (1985) described a framework in which both a decomposition of item and person parameters is implemented to model learning effects.
2. A model based on the decomposition of response probabilities in terms of more elementary probabilities. The most prominent example in this category is the multicomponent latent trait model (MLTM; Embretson, 1980, 1984), which was later extended into the conjunctive Rasch model (Maris, 1995). The basic principle of this model is that the product of response probabilities for the basic components determines the response probability for performance on the full item.

These two types of componential models can be used to model the full performance on items in terms of basic components. The models developed here, however, were for componential items—items derived from a decomposition of other items. They measure only partial aspects (components) of the full performance required on the original item. The MLTM models performance on total task items, and requires observations on total tasks as well as subtasks. The models described here model performance on the subtasks themselves, without specifying a relationship between the observations on these subtasks to observations on the corresponding original total task, and without requiring observations on the corresponding total tasks.

If the additional assumption is made, however, that the probabilities of total task responses can be derived from the probabilities of subtask responses, with the probability of success on the total task equal to the product of the success probabilities for the subtasks, then the MLTM corresponds to Model 1. The probability of having all components correct is simply the product of the probabilities of having each separate component correct. Model 1 assumes local item independence; it then follows that, for the synonym task data example, if the total task probabilities behave in correspondence with the componential subtask response patterns, the MLTM is rejected by the data. This is because there are strong indications of local item dependence. To adequately test the MLTM for the synonym task data, however, data for total task items are also required. Janssen & De Boeck (1997) found that, using total task data, the MLTM had to be rejected.

The models described here are not designed to relate componential items to global items, but to relate componential items to each other by specifying the kind of dependencies they show. These models, however, are closer to the second type of componential models (e.g., the MLTM) than to the first type (e.g., the linear logistic test model). By adding the postulate that the probability of having a global item correct equals the probability of a joint success on all componential items, a decomposition of the response probability can be obtained effectively for the global item in terms of more elementary probabilities.

Jannarone's (1986) conjunctive kernel model is similar to the models presented here. It can also be used when observations are available for componential items only, and it allows for local item

dependence. Jannarone's model, however, is unidimensional; furthermore, it restricts the weights for the person parameter in case of a joint success on two items to be equal to 3, based on the person parameter for success on each separate item and the person parameter for the conjunction of successes. The models described here are a generalization of Jannarone's conjunctive kernel model. The distinctions among the various componential models are important for studying cognitive components; they refer to different ways that components interact and, thus, to the type of model that can be applied to explain total task performance in terms of performance on underlying components.

References

- Ackerman, T. A., & Spray, J. A. (1986). *A general model for item dependency* (ACT Research Report 87-9). Iowa City IA: ACT.
- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27-41). Amsterdam: North Holland.
- Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245-275). New York: Academic Press.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading MA: Addison-Wesley.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge MA: MIT Press.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.
- Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Embretson (Whitley), S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.
- Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175-186.
- Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32, 277-294.
- Fischer, G. H. (1973a). Ein Beitrag zum Speed-Power Problem [A contribution to the speed-power problem]. In G. Reinert (Ed.), *Bericht ueber den 27. Kongress der Deutschen Gesellschaft fuer Psychologie in Kiel 1970*. Goettingen, Germany: Verlag fuer Psychologie.
- Fischer, G. H. (1973b). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131-155). New York: Springer-Verlag.
- Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement*, 6, 397-416.
- Geerts, G., & Heestermans, H. (Red.). (1984). *Van Dale Groot woordenboek der Nederlandse taal* [Van Dale extensive dictionary of the Dutch language] (11de uitgave). Utrecht/Antwerpen, The Netherlands: Van Dale Lexicografie.
- 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 dependence among test items. *Psychological Methods*, 2, 261-277.
- Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, 51, 357-373.
- Janssen, R. (1994). *Psychometric modeling of componentially designed synonym tasks*. Unpublished Dissertation, University of Leuven, Belgium.
- Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement*, 21, 37-50.
- Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, 20, 155-168.

- Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59, 149–176.
- Kelderman, H., & Steen, R. (1993). *LOGIMO* [Computer program]. Groningen, The Netherlands: ProGAMMA.
- Maris, E. (1995). Psychometric latent response models. *Psychometrika*, 60, 523–547.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195–215.
- Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, 61, 41–71.
- Patz, R. (1996). Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress (Doctoral Dissertation, Carnegie-Mellon University, Pittsburgh, 1997). *Dissertation Abstracts International*, 57/11, 7030.
- Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, 53, 349–359.
- Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169–194). New York: Academic Press.
- Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, 1, 81–97.
- Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human intelligence*. Hillsdale NJ: Erlbaum.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 104, 385–395.
- Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247–260.
- Verhelst, N. D., & Glas, C. A. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, 58, 395–415.
- 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. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12, 353–364.
- Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, 105, 276–289.
- Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60, 181–198.
- Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished Master's Dissertation, University of Melbourne, Australia.
- Wu, M. L., Adams, R., & Wilson, M. (1997). *CONQUEST* [Computer program]. Melbourne, Australia: Australian Council for Educational Research.
- Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 275–295). Hillsdale NJ: Erlbaum.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Acknowledgments

The research was completed while the first author was a research assistant for the National Fund for Scientific Research of Belgium.

Authors' Addresses

Send requests for reprints or further information to Machteld Hoskens, University of California at Berkeley, School of Education, 4623 Tolman Hall, Berkeley CA 94720, U.S.A.; or Paul De Boeck, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium, Europe. Email: hoskens@socrates.berkeley.edu; paul.deboeck@psy.kuleuven.ac.be.