JOURNAL OF APPLIED MEASUREMENT, 7(3), 335-350

Multicomponent Latent Trait Models for Complex Tasks

Susan E. Embretson
Georgia Institute of Technology

Xiangdong Yang
University of Kansas

Contemporary views on cognitive theory (e.g., Sternberg and Perez, 2005) regard typical measurement tasks, such as ability and achievement test items, multidimensional, rather than unidimensional. Assessing the levels and the sources of multidimensionality in an item domain is important for item selection as well as for item revision and development. In this paper, multicomponent latent trait models (MLTM) and traditional multidimensional item response theory models are described mathematically and compared for the nature of the dimensions that can be estimated. Then, some applications are presented to provide examples of MLTM. Last, practical estimation procedures are described, along with syntax, for the estimation of MLTM and a related model.

Requests for reprints should be sent to Susan E. Embretson, School of Psychology, Georgia Institute of Technology, 654 Cherry St., Atlanta, GA 30332-0170, USA, e-mail: susan.embretson@psych.gatech.edu.

Complex measurement tasks, such as ability and achievement test items, are usually multidimensional, rather than unidimensional, in nature. Contemporary views on cognitive theory regard even the most simple ability items (Sternberg and Perez, 2005), such as a vocabulary synonym task or a two-dimensional rotation, as involving multiple stages. In two-dimensional rotation tasks, for example, the goal is to find the figure that is a rotation of the figure in the stem. Multiple stages are involved in this simple task, including encoding the figures in the stem and alternatives, mentally rotating the stem figure and then comparing features of the stem figure to the alternatives. The difficulty of the various processing stages may vary between items and between persons, thus creating multidimensionality.

Assessing the levels and the sources of multidimensionality in an item domain is important not only for item selection, but also for item revision and item development. Both the dimensionality of the resulting measure (i.e., unidimensional versus multidimensional) and the nature of the construct that is measured depends on an adequate assessment of dimensionality in the item domain. As pointed out by Messick (1995), both construct-relevant and construct-irrelevant sources of variation may exist in the broader item domain.

The multidimensional normal ogive model (Bock, Gibbons and Muraki, 1988) or the multidimensional logistic model (Reckase and McKinley, 1991) can be applied to identify multiple dimensions in the data and to assess the dependence of each item on the dimensions. The approach is similar to a factor analysis of items; in fact, under certain conditions the IRT model and the factor model for binary data are identical (Takane and DeLeeuw, 1987). If the central dimension in the items is the target dimension, then only items with no significant discriminations on the smaller dimensions would be selected for the measure.

Although the traditional multidimensional IRT models described above, as well as factor analysis, are often applied to understand dimensionality in the item domain, they are somewhat limited in assessing complex cognitive tasks. First, since these models are exploratory, the nature of the central dimension is unclear. It could be that a smaller dimension, or another rotation of the dimensions, would be best to reflect construct-relevant sources of variation. Second, the mathematical relationships in the traditional multidimensional IRT models are not appropriate for assessing the multiple processing stages in complex items. In the multidimensional IRT models and the factor models, a compensatory relationship is specified between the dimensions. As elaborated below, a compensatory model does not properly reflect the sequential dependency among the processing stages and hence, does not assess adequately the sources of multidimensionality in the item domain.

This paper contains several sections. First, alternative IRT models are described mathematically and compared for how they consider the nature of multidimensionality in item domains. Specifically, the traditional compensatory multidimensional models are contrasted with the multicomponent latent trait model (MLTM) and a related model. Second, some applications are described to provide examples. Third, practical estimation procedures are described along with illustrative results that show their relative advantages. Fourth, syntax for the estimation of MLTM and a related model are presented in the Appendix.

Multidimensional IRT Models

Compensatory multidimensional models. In the compensatory models, the relative strength of the multiple dimensions in an item is indicated by the discrimination parameters. For example, in the multidimensional logistic model (Reckase and McKinley, 1991), the probability of an item response is given as follows:

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i)=\frac{\exp\left(\sum_m \alpha_{im}\theta_{sm}-\beta_i\right)}{1+\exp\left(\sum_m \alpha_{im}\theta_{sm}-\beta_i\right)}, \quad (1)$$

where X_{is} is the response of person s to item i, θ_{sm} is the trait level for person s on dimension m, β_i is the difficulty of item i and α_{im} is discrimina-

tion for item i on dimension m. In this model, the probability that a person passes an item depends on the difficulty of the item and a weighted combination of the multiple abilities, $\Sigma_m \alpha_{im} \theta_{sm}$. A low level on one trait can be compensated by an exceptionally high level on another trait for items with significant discriminations on more than one dimension.

The multidimensional normal ogive model (Bock, Gibbons and Muraki, 1988) is similar to the multidimensional logistic model but a different function relates the item response to the trait composite. That is, the model specifies a cumulative normal function, which produces an item characteristics curve which is very similar to the logistic model. One advantage of the normal ogive function, however, is that the resulting person and item estimates are directly related to binary factor analysis.

Multicomponent latent trait models (MLTM). Although the multidimensional item response models can be applied to complex tasks, they are not usually appropriate to assess the cognitive sources of multidimensionality. Cognitive models for tasks typically postulate a flow of information from one stage to another. The stages are sequentially dependent; correct processing on a later stage requires correct information from earlier stages. Thus, if a task depends on the joint outcome to several processing stages, a compensatory model is inappropriate. High trait levels for processing the later stages cannot compensate for low trait levels (and likely incorrect processing) on the earlier stages. A more appropriate model for multidimensionality would be based on a continued product of processing outcome probabilities, as follows:

$$P(X_{isT} = 1) = \prod_k P(X_{isk}), \qquad (2)$$

where $P(X_{isT} = 1)$ is the probability of success for person s on item i and $\Pi_k P(X_{isK})$ is the product of success on each processing component, k, given the correct outcome of the preceding component.

The multidimensional latent trait model (MLTM; Whitely[1], 1980) combines a continued product model of the response process as in Equation 2 with an IRT model to reflect individual differences in component trait levels. Thus, both component trait levels and component item difficulties are estimated. In MLTM, each component response probability, $P(X_{isk})$, depends on the difficulty of the component in the items and on the person's component trait level, according to a Rasch model, as follows:

$$P(X_{isT} = 1 \mid \theta_s, \beta_i) = \prod_k \frac{\exp(\theta_{sk} - \beta_{ik})}{1 + \exp(\theta_{sk} - \beta_{ik})}, \qquad (3)$$

where θ_{sk} is the trait level of person s on component k and β_{ik} is the difficulty of item i on component k. Notice that the right side of the equation contains Rasch models for the probability of success on each component. Other parameters, such as guessing, may be incorporated into MLTM as well (see Embretson, 1985).

A generalization of MLTM, the general component latent trait model (GLTM, Embretson, 1984), incorporates a mathematical model to relate component item difficulty to item content features. Like the linear logistic test model (LLTM, Fischer, 1973), item difficulty is given by the weighted sum of underlying stimulus factors, θ_{ikm}. In GLTM, the mathematical model is at the component level. For example, paragraph comprehension items, in which a short paragraph is followed by a question based on the paragraph, has two major components, text representation and decision (see Embretson and Wetzel, 1987 and Gorin and Embretson, in press). The difficulty of each component is related to stimulus features in the item; text representation depends on vocabulary level and syntactic complexity while decision depends on the inference level and the amount of relevant text for the question. For GLTM, scores on these variables for each item become part of the model. That is, component item difficulty, β'_{sk}, is predicted from the scored variables as follows:

$$\beta'_{ik} = \sum_m \eta_{km} q_{ikm} + \eta_0, \qquad (4)$$

where θ_{ikm} is the score of stimulus factor m on component k for item i, η_{km} is the weight of stimulus factor m in component k and η_0 is an inter-

cept. The model in Equation 4 is directly analogous to a regression model of component item difficulty in which the independent variables are the scored item features.

The full GLTM combines the MLTM for the relationship of the components to the total item response with the mathematical model for component item difficulty as follows:

$$P(X_{ir}=1|\theta_i,\eta_i)=\prod_k \frac{\exp\left(\theta_{ik}-\sum_m \eta_{km}q_{ikm}+\eta_0\right)}{1+\exp\left(\theta_{ik}-\sum_m \eta_{km}q_{ikm}+\eta_0\right)}. \quad (5)$$

Although estimates of stimulus features impact on item difficulty can be obtained by ordinary multiple regression, the standard errors will be much larger than those from the full information GLTM given in Equation 5.

Comparison of models. The dimensions that results from applying MLTM may differ substantially from the compensatory multidimensional IRT models that assess dimensionality. Perhaps most importantly, the relationship between theory and data differs between the compensatory multidimensional IRT models and MLTM. Although both types of models decompose items into underlying sources, the dimensions from the compensatory multidimensional IRT models may not represent processes for several reasons. First, the mathematical forms of the model are quite different. Processes are only sometimes regarded as compensatory in complex tasks, so the fit of the models to actual data could vary substantially. Second, the compensatory multidimensional IRT models may not reflect processing similarity because the models depend on item intercorrelations. Components with similar processing may not correlate highly if, for example, individuals do not really vary in processing competency or if the processes are confounded with other processes. On the other hand, separate processes may be substantially correlated in a particular population if they are linked by common patterns of experiences, educational prerequisites and even genetics. The result of such influences again implies that the processes will not correspond to separate dimensions.

In contrast, MLTM relies on theory to both identify and guide the operationalization of components in tasks. Thus, the dimensions identified in MLTM are theory-driven rather than empirically-driven. Of course, the theory could be wrong so that the dimensions that are identified may have little impact on task performance. Thus, methods to evaluate the theory are important in applications of MLTM.

Illustrative Examples of Applications

Applications of MLTM and GLTM models have three basic uses. First, the theoretical model of task processing, which is reflected in the component outcomes, may be tested for adequacy. Fit may be evaluated for the linkage of the components to the total item to test the theory. Second, estimates of the difficulty of the components within each item may be obtained. Item component estimates are useful in guiding test development for the construct representation aspect of construct validity; that is, how to select items to reflect target sources of cognitive complexity. Third, estimates of person ability on each component may be obtained. Component ability estimates may be useful in understanding the nomothetic span aspect of construct validity; that is, how the components are related to external measures and criteria.

Two types of applications have appeared in the literature; studies in which both component responses and total item responses are available and studies in which only the total item response is available. Examples of both types will be described here.

Responses to both components and total tasks. MLTM and GLTM (Whitely[1], 1980; Embretson, 1984) were developed for data in which both component responses and total item responses are assessed. Two different designs have been used to obtain both component and total responses. Both methods involve constructing partial tasks to assess component outcomes.

In the first design, the component responses and the total item responses are observed on the same item. For example, Whitely[1] (1980) administered standard verbal analogy items, thus observing the total item response. Then, some time later, the component outcomes were assessed

by tasks constructed from the same items; image construction and response evaluation. For the image construction task, the person was presented an analogy stem (with no response alternatives) and asked to specify the rule governing the analogy. Then, for the response evaluation component, the person was given a correct rule along with the item stem and asked to identify the correct response alternative. Both tasks can be scored for accuracy; thus, joint response patterns are available for each item. An estimation procedure developed for joint response patterns (Embretson, 1984) can be applied.

Table 1 shows joint responses for two components and the total item, as well as the MLTM terms that correspond to the pattern. When response patterns are collected in this manner, two additional parameters can be estimated for MLTM to link the component outcomes to the total probabilities. In Table 1, a is the probability of passing an item when the components are passed and g is the probability of passing the item when one or more components are failed.

In the second design, the component responses and the total item response are assessed on different items. In this case, the full data set is modeled simultaneously (see description of estimation below) as one long vector of responses. For example, Maris (1995) applied a special algorithm for missing data to Janssen and DeBoeck's (1997) data to estimate two components that were postulated to underlie success on synonym items. These components were 1) generation of a potential synonym and 2) evaluation of the potential synonym. The resulting estimation of MLTM parameters had two implications for construct validity. First, the cognitive model of synonym items was further supported by the results. The two-component model had good fit to the synonym item responses, thus supporting the importance of both components. Further, the nature of the task was elaborated by comparing the relative strength of the components; success on the generation component was far more crucial to item solving. Second, a related study (Janssen, DeBoëck, and Van der Steene, 1996) indicated that the relative strength of the generation and evaluation components in synonym items influenced the correlation of the persons' performance on the synonym test with other tasks.

Although Maris (1995) does not present individual item data, a scatterplot of component difficulties would probably be quite similar to Figure 1, which shows uncorrelated component difficulties. Notice that items in which difficulty stems primarily from only one component can be identified. That is, such items are so easy on one component that item difficulty depends only on their differences on the other component. Construct validity can be controlled by selecting items to represent primarily one component or a combination of components, as desired.

Total item only. Obtaining component item responses requires a specially designed study as test data typically does not include component item responses. Thus, it is desirable to be able to extract component data from the total task response alone. Such a possibility is obviously

Table 1

Component and Total Response Pattern Data for Two Components.

	Components		Total	
Pattern	C1	C2	T	Model Probability
P1	1	1	1	$aP_{C1}P_{C2}$
P2	1	1	0	$(1-a)\,P_{C1}P_{C2}$
P3	0	1	1	$g\,(1-P_{C1})P_{C2}$
P4	0	1	0	$(1-g)\,(1-P_{C1})P_{C2}$
P5	1	0	1	$g\,P_{C1}(1-P_{C2})$
P6	1	0	0	$(1-g)\,P_{C1}(1-P_{C2})$
P7	0	0	1	$g(1-P_{C1})(1-P_{C2})$
P8	0	0	0	$(1-g)(1-P_{C1})(1-P_{C2})$

more exploratory in nature than the component response method, since the nature of the components is not defined *a priori*. Although the developments here are more pioneering, under certain circumstances estimates have been obtained from the total task alone. However, two developments in estimation methods, 1) an EM algorithm (Maris, 1995) and 2) a new quasi-Bayesian approach (Yang and Embretson, 2004) have lead to successful component estimation. The latter will be described below.

Maris' (1995) algorithm has been successfully applied to a constrained version of GLTM to permit component abilities to be estimated without subtasks. The component parameters can be constrained when separate studies have developed mathematical models or prior theory that leads to strong prediction of item difficulty. Embretson (1995) applied the Maris (1995) algorithm using GLTM to estimate individual differences in two underlying components that were postulated for abstract reasoning, working memory capacity and general control processes. The separate component abilities were identified by the underlying model for each component. For the Abstract Reasoning Test (ART), which consists of matrix problems, previous studies had shown that two variables that impact working memory, the number and the level of rules in the items, were strong predictors of item difficulty. General control processes, on the other hand, often are assumed equally difficulty across items. Figure 2 is a scatterplot of two component abilities, working memory capacity and general control processes that were extracted from standard ART items using GLTM (Embretson, 1995). Differing correlations of these two abilities with reference tests further supported their validity (see Embretson, 1995).

A similar approach was applied to identify underlying components in an aging study on spatial ability, which typically shows substantial age-related decline. Embretson and McCollam (2000) applied the Maris (1995) algorithm to GLTM to measure individual differences in both working memory capacity and general control processes. They found substantial age-related decline in both components. Such results are interesting because general control processing may be more amendable to intervention than working memory capacity.

Estimation of MLTM and GLTM

Total and Component Responses

Response pattern formulation of model. MLTM was formulated as a model of joint response patterns (Embretson, 1984) for binary test items. As described above, Table 1 shows the eight possible response patterns when the total item response and two component item responses are measured. To estimate the model parameters

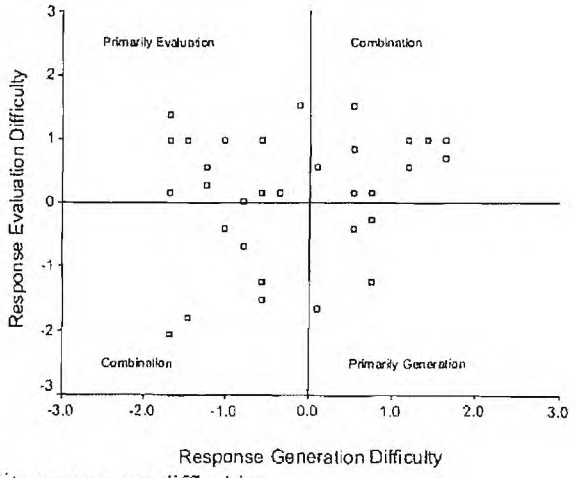

Figure 1. Scatterplot of item component difficulties.

from the joint response patterns, each possible pattern must be linked to terms in the model. Thus, the various patterns depend on the joint probabilities of the components, P_{c1} and P_{c2}, and terms to reflect the linkage of the component to the total. The term a is the probability of solving the total item when the components are solved, while g is the probability of solving the total item when one or more components are failed.

The MLTM for the joint response pattern of the component responses, X_{is}, and the total response, X_{isT} for person s on item i is given as follows:

$$P(\mathbf{X}_{is}, X_{isT}) = \left[g^{X_T} (1-g)^{1-X_T} \right]^{1-\prod X_k} \left[a^{X_T} (1-a)^{1-X_T} \right]^{\prod X_k} \left[\prod P_k^{X_k} Q_k^{1-X_k} \right], \quad (6)$$

where x_k is the response to component k, x_T is the response to the total item and P_k is the IRT model probability for component k, all of which are defined for the individual response pattern of person s on item i. The terms a and g are defined as above. In Equation 6, the exponents control the entry of terms into the probability of a specific response pattern.

Embretson (1984) presents proofs that show how the parameters may be estimated separately from each component and then linked to the total item response through response pattern frequencies. In this formulation, MLTM component parameters can be estimated with readily available software for the Rasch model, such as WINSTEPS or RUMM or even as an option in BILOG-MG. The linkage parameters, a and g, can be estimated by the relative frequencies of response patterns, over persons and items, as follows using the information in Table 2:

$$a = f_{P1} / (f_{P1} + f_{P2}), \text{ and} \quad (7)$$

$$g = (f_{P3} + f_{P5} + f_{P7}) / (f_{P3} + f_{P5} + f_{P7} + f_{P4} + f_{P6} + f_{P8}).$$

To demonstrate the model, item response data were simulated so that the true parameters are known. Data were simulated for 15 items and 2,000 persons. The component abilities were obtained as random samples from a bivariate normal distribution with means of zero, variances of 1 and an intercorrelation of zero. The item parameters were selected to yield total item probabilities in the range of .4 to .6 and to have moderate to strong differences in component difficulty, so that the impact of varying combinations of ability will be distinguishable. Item responses were then generated by computing the modeled probability to each item for each person and comparing the result to a random draw from a uniform probability distribution ranging from .00 to 1.00. If the computed probability exceeded the random draw, the component was passed; otherwise, the component was failed. Total item probabilities were computed as

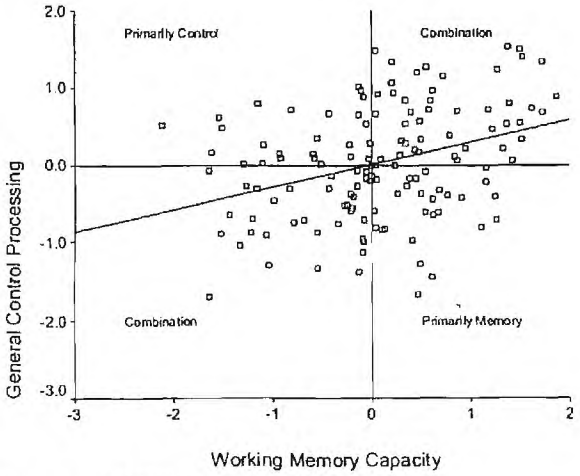

Figure 2. Scatterplot of component abilities.

the product of the component probabilities and similarly compared to a random draw from a uniform probability distribution.

Component item parameters were estimated by applying DeBoeck and Wilson's (2004, pp. 68-71) formulation of the Rasch model as a nonlinear mixed model, using the SAS NLMIXED program. Although a standard Rasch model program could be applied, NLMIXED allows more flexibility in models and estimation procedures, which is important for estimating MLTM from the total item response only, as discussed below.

In the Appendix, SAS syntax is shown for the "Single Component Rasch Model", which is applied to item response data for each component separately. The data must be structured as a single response vector, with each item on a separate line for each person. Also, dummy variables, one for each item, must be included on each line to indicate the item.

Since MLTM is a special type of IRT model, it is important to demonstrate the adequacy of the estimation procedures. Figure 3 shows the recovery of the component item difficulties from the simulated data. It can be seen that the regression of the estimates on the true parameter values yielded a very high squared multiple correlation ($r^2 = .9977$), with the regression coefficients ($b = 1.007$, $a = -.025$) close to a perfect scaling of 1 and 0, respectively, as expected.

The MLTM a and g parameters were estimated from the response pattern frequencies in Table 2 as follows:

$$a = 8610 / (8610 + 4910) = .637, \text{ and}$$

$$g = (2110 + 2241 + 353) / (2110 + 2241 + 353 + 5184 + 5170 + 1311) = .287.$$

Since the average product of the components over items was .64, the a parameter accurately reflects the probabilistic method by which the product of the component probabilities was related to the total task outcome.

Table 3 shows the correlation of the estimated component abilities with the true abilities under the joint response pattern method, which is labeled as "Component Only". The correlations are quite high (.906), as expected.

Thus, the estimation procedures adequately recover the (known) parameters. In real data, it is important to evaluate the fit of the model to both the component subtask responses and to the total item. For the component subtask data, the adequacy of the IRT model (i.e., Rasch model) can be evaluated by standard methods for single dimensions. Hence, these will not be elaborated here. However, the predictability of the total item response from the component estimates is a unique concern in applications of MLTM and will be elaborated here.

The fit of the component model to the total item can be determined in two different ways (see

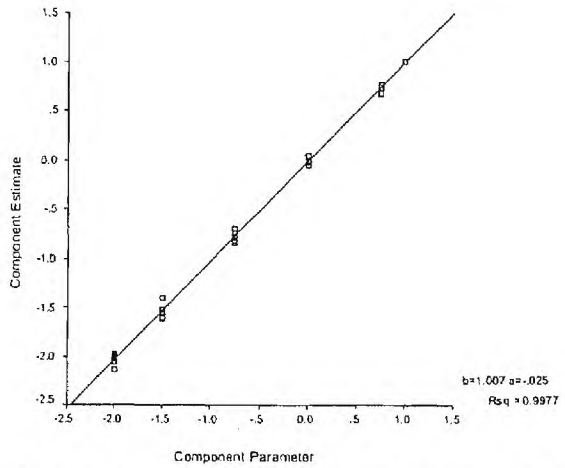

Figure 3. Item Parameter Recovery from MLTM: Joint Response Pattern Method

Embretson, 1984). First, a goodness of fit test can be performed for each item. That is, the frequency of responses across score groups varying in total score is compared to the frequency predicted by the component model. Second, observed total item probabilities can be compared to predictions from the component model. In this case, expected and observed probabilities are correlated for persons and for items.

Simultaneous formulation of the model. In this formulation of the model, the component parameters are estimated simultaneously rather than as joint response patterns. This method requires a special data structure. The complete component and total item response data are treated as a single response vector, such that the response to each task or subtask for each item for each person is a separate line. That is, if an item has two component tasks, then three lines are required for the responses to the total item and the two component subtasks.

Table 4 shows required data structure. In Table 4, each item has a total task response and two subtask responses, which are nested under each subject. Y is the response given, while $X1$ and $X2$ indicate the involvement of the components. For example, if $X1 = 1$ and $X2 = 0$, then the response in Y is to the subtask for the first component. If both $X1 = 1$ and $X2 = 1$, then the response in Y is for the total task. Dummy variables are further needed to indicate the item. For example, for each subject, Item1 is equal to 1 exactly three times, to designate the total task response and the two component subtask responses for Item 1. Standard program packages, such as SPSS, have a menu- driven algorithm for transforming data to this structure.

The single response vector, denoted as y, includes the response x_{isk} for each person s, each item i on all components (and total) k. To estimate component parameters, dummy variables, c_{ik}, are needed to define data type (i.e., Compo-

Table 2

Descriptive Statistics for MLTM Response Patterns

Pattern	Responses	Frequency
1	111	8610
2	110	4910
3	011	2110
4	010	5184
5	101	2241
6	100	5179
7	001	353
8	000	1311

Note: Each pattern describes responses to two components and total, respectively.

Table 3

Correlations of True and Estimated Trait Levels Under Different Estimation Methods

	Theta1 Parameter	Theta2 Parameter
Theta1 Parameter	1	.014
Theta2 Parameter	.014	1
Theta1 Simultaneous	.933	.021
Theta2 Simultaneous	.024	.934
Theta1 Component Only	.906	.002
Theta2 Component Only	.016	.906
Theta1 Quasi-Bayesian	.821	.187
Theta2 Quasi-Bayesian	.078	.816

nent1, Component2, etc.). The total item response would be indicated by "1" for all c_k. Thus, MLTM is formulated as follows:

$$P(X_{isk}) = \prod_k \left[P_k^{y} Q_k^{1-y} \right]^{c_k}. \quad (8)$$

In this formulation of the model, the c_k exponents determine whether or not a probability from component k is relevant to the response y. A Rasch model underlies each component probability, P_k, as indicated in Equation 3. A section in the Appendix shows the NLMIXED syntax for the simultaneous estimation method for MLTM in the section, "MLTM: Simultaneous Component and Total".

Figure 4 shows the recovery of item parameters from the simultaneous estimation method. As with the joint response pattern estimation procedure the regression of the estimates on the true parameter values yielded a very high squared multiple correlation ($r^2 = .9930$), with the regression coefficients ($b = 1.057$, $a = -.031$) close to a perfect scaling of 1 and 0, respectively, as expected.

Table 3 shows the correlation of the estimated component abilities with the true abilities under the simultaneous method. It can be seen that the correlation is somewhat higher than the joint response pattern method for both components. This increased correlation represents the increased information provided by including the total item response in the simultaneous solution.

Figure 5 shows the standard errors for each ability level on Component 1 under the two different estimation procedures, component only estimates (i.e., from the joint response pattern method) and simultaneous solution estimates. It can be seen that for all abilities, the simultaneous solution lead to generally lower standard errors. Again, this reflects the increased information provided by the total response for estimating the component ability lowers the standard errors.

As described above for the joint response estimation procedure, fit may be assessed for both the components and for the total response. These methods will not be repeated here, as they are shown in Embretson (1984).

Table 4

Sample Data Setup for Simultaneous Model Estimation: Four Items and Two Persons

		Y	X1	X2	ITEM1	ITEM2	ITEM3	ITEM4
ID	1	0	1	1	1	0	0	0
		0	1	1	0	1	0	0
		0	1	1	0	0	1	0
		0	1	1	0	0	0	1
		1	1	0	1	0	0	0
		1	1	0	0	1	0	0
		1	1	0	0	0	1	0
		1	1	0	0	0	0	1
		0	0	1	1	0	0	0
		0	0	1	0	1	0	0
		0	0	1	0	0	1	0
		0	0	1	0	0	0	1
	2	0	1	1	1	0	0	0
		0	1	1	0	1	0	0
		0	1	1	0	0	1	0
		0	1	1	0	0	0	1
		1	1	0	1	0	0	0
		1	1	0	0	1	0	0
		1	1	0	0	0	1	0
		1	1	0	0	0	0	1
		0	0	1	1	0	0	0
		0	0	1	0	1	0	0
		0	0	1	0	0	1	0
		0	0	1	0	0	0	1

Simultaneous estimation of GLTM. With GLTM, the scored features of items replace component item difficulty in the model. Thus, the model for the simultaneous formulation of the GLTM is identical to Equation 8, except that the component item probability, P_k, is given by a weighted combination of item features, as shown in Equation 5. To estimate GLTM for data on 15 items, the data file must contain scores for items on stimulus features that are related to component or total difficulty. In the Appendix, the code given under "Single Component LLTM" can be used to estimate the weights for the scored stimulus features in a single component with two scored stimulus features, *q1* and *q2*. For GLTM, the code for GLTM: Simultaneous Component and Total Estimation reflects the replacement of the dummy variables for items with the scored stimulus features for each component.

To illustrate the model on the 15-item simulation data set, two scores for stimulus features were simulated for each component. The multiple correlation of the scores with true component item difficulty was .751 and .647, respec-

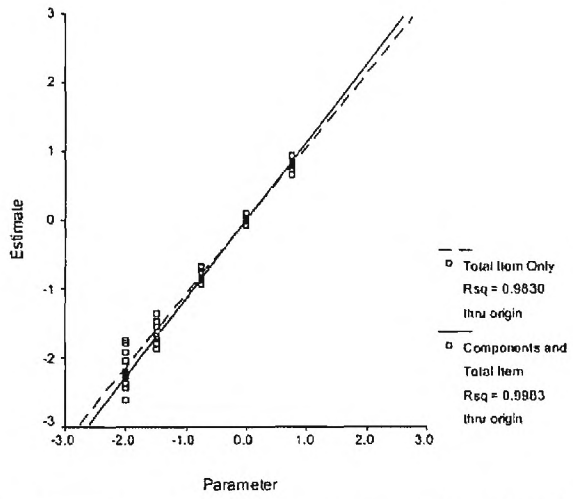

Figure 4. Item Parameter Recovery from MLTM: Simultaneous and Quasi-Bayesian Methods.

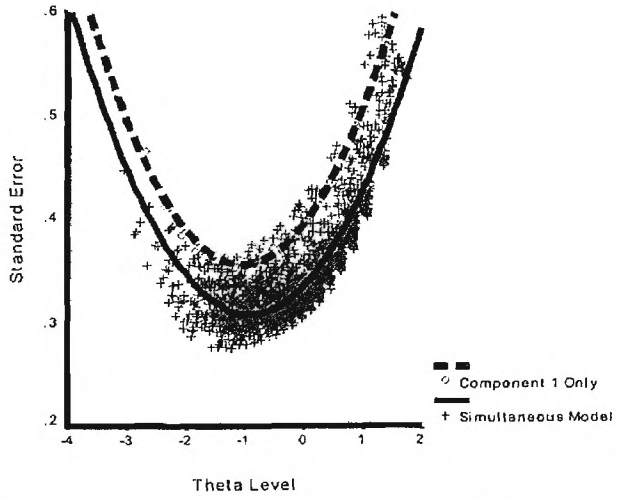

Figure 5. Ability Measurement Erors: Single Component versus Simultaneous Component and Total Item.

tively, for Component 1 and Component 2. With GLTM, the quality of the mathematical model for the data is evaluated at the component level. Table 5 shows tests for goodness of fit for each component. To evaluate model fit (see Embretson, 1997), three alternative GLTM's were specified for each component as follows: 1) a null model, in which all items are assumed equal in component difficulty, 2) the cognitive model, in which the component item difficulty is predicted by the scored features and 3) a saturated model, in which each item has an estimated difficulty on each component (i.e., the Rasch models in MLTM). A goodness of fit statistic to determine if the cognitive model provides significant prediction is –2 times the difference in the log likelihood from the null model to the cognitive model. This statistic is approximately distributed as χ^2 with degrees of freedom equal to the change in the number of parameters. For example, a constant component item difficulty is estimated for the null model while the weights for two content variables and a constant are estimated for the cognitive model. Thus, the difference in the number of estimated parameters is two. Evaluating the significance of this fit statistic is analogous to evaluating the significance of the regression of component item difficulty on the scored content features. The cognitive model yields significant prediction for Component 1 ($\chi^2 = 5136$, $df = 2$, $p < .001$) as well as for Component 2 ($\chi^2 = 3203$, $df = 2$, $p < .001$).

A second significance test can be applied to evaluate the significance of information about item difficulty that is not predicted by the model. That is, the cognitive model is compared to the saturated model (i.e., the Rasch model). It can be seen in Table 5 that the change in the number of parameters is 12 and that the goodness of fit values are quite large for both Component 1 ($\chi^2 = 3728$, $df = 12$, $p < .001$) as well as for Component 2 ($\chi^2 = 5688$, $df = 12$, $p < .001$). These tests indicate that the cognitive model does not capture all the differences in component difficulty between the items.

Last, a fit index (Embretson, 1997) can be computed, based on the pattern of likelihoods between the three models for each component. That is, the fit index h, is given as follows:

$$h = [-2 \ln L_{null} - (-2 \ln L_{model})]/$$
$$[-2 \ln L_{null} - (-2 \ln L_{saturated})]^{1/2}. \quad (9)$$

For the data in Table 5, the fit index for Component 1 is computed as follows:

$$\eta = [116964 - 111828)]/$$
$$[116964 - 108090]^{1/2} = .762.$$

This value is close to the multiple correlation of .751 of the true component item difficulties with the scored variables. The fit index for Component 2 was computed as .600 which is reasonably close to the multiple correlation of .647 for the true parameters.

Table 6 shows the GLTM parameter estimates and standard errors for the scored features on each component. It can be seen that the parameter estimates are large relative to their small standard errors; thus, all parameter estimates were statistically significant.

Also in Table 6 are estimates of the component ability variances. The estimates were substantially less than 1, which was the true generating value. These lower estimates result from the predicted component item difficulties having a more restricted range of values than the true item parameters. This is analogous to a regression effect in predicting item difficulties.

Table 5

Fit Statistics for Alternative GLTMs of Simulated Data.

	Component 1			Component 2		
Model	–2lnL	Δdf	$\Delta\chi^2$	–2lnL	Δdf	$\Delta\chi^2$
Null	116964	.....	.....	117179	....	....
Cognitive	111828	2	5136	113976	2	3203
Saturated	108090	12	3728	108288	12	5688
Fit Index	.762				.600	

Total Item Only

Typical test data does not include component item responses. If components can be extracted from the total item response only, the utility of MLTM and GLTM would be greatly increased. Previous studies (Embretson, 1995; Embretson and McCollam, 2000) were able to extract some component information (i.e., the person abilities) from the total task alone. These studies applied the Maris (1995) algorithm to estimate component abilities by constraining the component difficulties in the items. This is a restrictive set of conditions for practical application. Further, the program that was used, COLORA, is no longer readily available.

Yang and Embretson (2004) have developed an initial version of a quasi-Bayesian estimation procedure for the components, using the total item response only. Their procedure is implemented through the SAS NLMIXED program. Since the procedure is new and needs further development for practical application, the syntax is not given in the Appendix. However, results on the simulated data set will be presented to illustrate the degree of effectiveness that can be obtained from the procedure.

Figure 4 shows the regression of estimated component item difficulties on the true parameters for both the total item and component method (simultaneous method above) and the quasi-Bayesian method. It can be seen that the squared multiple correlation for the quasi-Bayesian procedures ($r^2 = .9830$) is nearly as good as for the total item and component outcomes together ($r^2 = .9983$). Table 3 presents the correlation of the quasi-Bayesian person estimates with the true generating abilities. It can be seen that although the correlations are lower than either case when component subtasks are available, the correlations are still above .80. Thus, the feasibility of the quasi-Bayesian procedure is demonstrated. Further developments are needed to assure its practical scope.

Summary

Multidimensionality is an inherent aspect of cognitive measures that employ complex items to measure ability. Traditional compensatory multidimensional IRT models were reviewed and shown to not adequately represent contemporary views of the latent sources of multidimensionality. That is, complex items involve multiple processing stages, each of which must be completed correctly to solve the item. The compensatory feature in the traditional multidimensional models would allow low ability for one stage to be compensated by high ability for another stage. Conceptually, however, an unrelated ability should have no impact on completing a given stage. Thus, multicomponent latent trait models, MLTM and GLTM, were developed to more adequately represent the underlying sources of multidimensionality.

In this paper, MLTM and GLTM are described mathematically and compared to traditional multidimensional IRT models. Applications of MLTM and GLTM have three important implications: 1) the fit of the postulated cognitive model may be assessed, thus contributing to construct validity for the measure, 2) item differences in component difficulty may be estimated,

Table 6

GLTM Component Item Parameter Estimates from NLMIXED.

Variable	Parameter Estimate	Standard Error	*DF*	*t*-Value	Prob
Component1 [–2logL=111828]					
Predictor 1 (c1b1)	0.5323	0.01124	1999	47.36	<.0001
Predictor 2 (c1b2)	0.4045	0.01236	1999	32.73	<.0001
Constant (d)	–0.4568	0.02140	1999	–21.35	<.0001
Person Var. (v1)	0.6920	0.03248	1999	21.31	<.0001
Component2 [–2logL=113976]					
Predictor 1 (c2b1)	0.2915	0.01273	1999	22.90	<.0001
Predictor 2 (c2b2)	0.4862	0.01300	1999	37.41	<.0001
Constant (d)	–0.3106	0.01994	1999	–15.58	<.0001
Person Var. (v2)	0.5505	0.02623	1999	20.98	<.0001

which permits items to be selected for targeted sources of cognitive complexity and 3) individual differences in component difficulty may be estimated, which permits the differential validity of the underlying abilities for predicting external measures to be assessed.

Examples of applications and estimation procedures are described for both MLTM and GLTM. A new estimation algorithm, which permits estimation of component levels from the total item alone, is described. This procedure greatly increases the utility of both MLTM and GLTM, as they can be applied to standard item response data. More research is needed to make this procedure more practically available.

Footnote

[1] Susan Embretson has also published as Susan Whitely.

References

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

DeBoeck, P., and Wilson, M. (2004). *Explanatory item response models*. New York: Springer-Verlag.

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

Embretson, S. E. (1985). *Test design: Developments in psychology andpsychometrics*. New York: Academic Press.

Embretson, S. E. (1995). Working memory capacity versus general central processes in intelligence. *Intelligence, 20*, 169-189.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden and R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-322). New York: Springer-Verlag.

Embretson, S. E., and McCollam, K. M. (2000). A multicomponent Rasch model for covert processes. In M. Wilson and G. Engelhard (Eds.), *Objective measurement: Theory into practice Vol. V* (pp. 203-218). Norwood, NJ: Ablex.

Embretson, S. E., and Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*, 175-193.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Gorin, J., and Embretson, S. (in press). Predicting item properties without tryout: Cognitive modeling of paragraph comprehension items. *Applied Psychological Measurement.*

Janssen, R., and De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 27*, 37-50.

Janssen, R., DeBoeck, P., and Van der Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence, 22*, 291-310.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *American Psychologist, 9*, 741-749.

Reckase, M., and McKinley, R. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 21*, 25-36.

Sternberg, R., and Perez, M. (2005). *Cognition and intelligence: Identifying the mechanisms of the mind.* New York: Cambridge University Press.

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

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

Yang, X., and Embretson, S. E. (2004). Some New Estimation Procedures for the Multicomponent Latent Trait Model. Unpublished manuscript. University of Kansas.

Appendix

SAS Syntax for PROC NLMIXED

Single Component Rasch Model

```
/*component 1 only*/
proc nlmixed data =jam.mltm12 noad;
parms c1b1-c1b12=0
          v1=1;
b1par = c1b1*item1+c1b2*item2+c1b3*item3+c1b4*item4+c1b5*item5+c1b6*item6
+c1b7*item7+c1b8*item8+c1b9*item9+c1b10*item10+c1b11*item11+c1b12*item12;
eta1 =th1-b1par;
eta =x1*eta1;
p =exp(eta)/(1+exp(eta));
model y ~binomial(1,p);
random th1 ~normal(0,v1)
           subject =id;
run;
```

Definitions. To estimate parameters for items, dummy variables must be created for each item: item1, item2,.....item15. To define values for trait level, a unique identifier must be assigned to each person. In the syntax above, "id" is the indicator variable for persons.

Single Component LLTM

```
/*component 1 only lltm*/
proc nlmixed data =jam.lltmdata2 noad;
parms c1b1-c1b2=0 d=0
          v1=1;
b1par = c1b1*q1+c1b2*q2+d;
eta1 =th1-b1par;
eta =x1*eta1*1.7;
p =exp(eta)/(1+exp(eta));
model y ~binomial(1,p);
random th1 ~normal(0,v1)
           subject =id;
run;
```

Definitions. To estimate parameters for the scored stimulus features, the data set must include scores, *q1*, *q2*....*qK*, on each line. To define values for trait level, a unique identifier must be assigned to each person. In the syntax above, "id" is the indicator variable for persons.

MLTM Simultaneous Component and Total

```
/*Simultaneous solution*/
proc freq data =jam.mltm5r;
   tables index1 / missprint;
   title '1-WAY FREQUENCY TABLE WITH MISSPRINT OPTION';
proc nlmixed data =jam.mltm5r noad qpoints=5;
parms c1b1-c1b15 = 0   /*difficulty parameters for component 1*/
c2b1-c2b15 = 0         /*difficulty parameters for component 2*/
a1=1   a2 =1;          /*variances of thetas for component 1 and 2*/
b1par=c1b1*item1+c1b2*item2+c1b3*item3+c1b4*item4+c1b5*item5+c1b6*item6
+c1b7*item7+c1b8*item8+c1b9*item9+c1b10*item10+c1b11*item11+
c1b12*item12+c1b13*item13+c1b14*item14+c1b15*item15;
b2par=c2b1*item1+c2b2*item2+c2b3*item3+c2b4*item4+c2b5*item5+c2b6*item6
+c2b7*item7+c2b8*item8+c2b9*item9+c2b10*item10+c2b11*item11+
c2b12*item12+c2b13*item13+c2b14*item14+c2b15*item15;
eta1 =1.7*a1*(th1-b1par);      /*th1: theta on component 1*/
```

(*Appendix continued on next page*)

(Appendix continued from previous page)

```
eta2 =1.7*a2*(th2-b2par);       /*th2: theta on component 2*/
eta =x1*eta1+x2*eta2;
deno =(1+x1*exp(eta1))*(1+x2*exp(eta2));
p =exp(eta)/deno;
model transl ~binomial(1,p);
random th1 th2 ~normal([0,0],[1,0,1])
          subject =id out=mltm5sco;
run;
```

GLTM Simultaneous Component and Total

```
/*Simultaneous solution*/
proc freq data =jam.mltm5r;
   tables indexl / missprint;
   title '1-WAY FREQUENCY TABLE WITH MISSPRINT OPTION';
proc nlmixed data =jam.mltm5r noad qpoints=5;
parms c1b1-c1b2 = 0   /*stimulus weight parameters for component 1*/
c2b1-c2b2 = 0         /*stimulus weight parameters for component 2*/
a1=1   a2 =1;          /*variances of thetas for component 1 and 2*/
b1par=c1b1*q1+c1b2*q2;
b2par=c2b1*q1+c2b2*q2;
eta1 =1.7*a1*(th1-b1par);       /*th1: theta on component 1*/
eta2 =1.7*a2*(th2-b2par);       /*th2: theta on component 2*/
eta =x1*eta1+x2*eta2;
deno =(1+x1*exp(eta1))*(1+x2*exp(eta2));
p =exp(eta)/deno;
model transl ~binomial(1,p);
random th1 th2 ~normal([0,0],[1,0,1])
          subject =id out=mltm5sco;
run;
```