Psychology Science Quarterly, Volume 50, 2008 (3), pp. 328-344

Understanding and quantifying cognitive complexity level in mathematical problem solving items

SUSAN E. EMBRETSON[1] & ROBERT C. DANIEL

Abstract

The linear logistic test model (LLTM; Fischer, 1973) has been applied to a wide variety of new tests. When the LLTM application involves item complexity variables that are both theoretically interesting and empirically supported, several advantages can result. These advantages include elaborating construct validity at the item level, defining variables for test design, predicting parameters of new items, item banking by sources of complexity and providing a basis for item design and item generation. However, despite the many advantages of applying LLTM to test items, it has been applied less often to understand the sources of complexity for large-scale operational test items. Instead, previously calibrated item parameters are modeled using regression techniques because raw item response data often cannot be made available. In the current study, both LLTM and regression modeling are applied to mathematical problem solving items from a widely used test. The findings from the two methods are compared and contrasted for their implications for continued development of ability and achievement tests based on mathematical problem solving items.

Key words: Mathematical reasoning, LLTM, item design, mathematical problem solving

[1] Dr. Susan Embretson, School of Psychology, Georgia Institute of Technology, 654 Cherry Street, Atlanta, GA 30332-0170, USA; email: susan.embretson@psych.gatech.edu

Since its introduction in 1973, the linear logistic test model (LLTM; Fischer, 1973) has been applied widely to understand sources of item complexity in research on new measures (e.g., Embretson, 1999; Gorin, 2005; Hornke & Habon, 1986; Janssen, Schepers, & Peres, 2004; Spada & McGaw, 1985). These applications require not only estimation of LLTM parameters on item response data, but a system of variables that represent theoretically interesting sources of item complexity. Advantages of item complexity modeling with LLTM include elaborating construct validity at the item level, defining variables for item design, predicting parameters of new items, item banking by sources of complexity and providing a basis for item design and item generation. These advantages will be reviewed more completely below.

However, despite the many advantages of applying LLTM to test items, it has been applied less often to understand the sources of complexity for operational test items. Instead, item difficulty statistics, such as classical test theory *p-values* or item response theory *b-values*, are modeled from item complexity factors using regression techniques. Often this approach is used because raw item response data cannot be made available. For example, Newstead *et al* (2006) modeled item difficulty of complex logical reasoning items from the Graduate Record Examination (GRE). Similarly, Gorin and Embretson (2006) modeled item parameters for verbal comprehension items from the GRE while Embretson and Gorin (2001) modeled item parameters for the Assembling Objects Test from the Armed Services Vocational Aptitude Battery (ASVAB). To date, item difficulty modeling has provided validity evidence for many tests, including English language assessments, such as TOEFL (Freedle & Kostin, 1996, 1993; Sheehan & Ginther, 2001) and the GRE (Enright, Morley, & Sheehan, 1999; Gorin & Embretson, 2006).

Unfortunately, the studies of item difficulty based on regression modeling have less clear interpretations about the relative impact of the complexity variables in the model. That is, the standard errors often are large since the regression modeling is applied to item statistics rather than raw item response data. Furthermore, the parameters estimated for the impact of the complexity variables are not useful for item banking because the usual properties of consistency and unbiasedness do not extend to the modeling of item statistics.

In this paper, applications of LLTM to items from a widely used test of mathematical reasoning will be contrasted with regression modeling of item statistics. The item complexity variables are based on a cognitive theory of mathematical problem solving that was developed for complex items. Prior to presenting the studies on mathematical reasoning, the LLTM and its advantages will be elaborated.

LLTM and related models

Several IRT models have been developed to link the substantive features of items to item difficulty and other item parameters. Unidimensional IRT models with this feature includes the Linear Logistic Test Model (LLTM; Fischer, 1973), the 2PL-Constrainted Model (Embretson, 1999) and the Random Effects Linear Logistic Test Model (LLTM-R; (Janssen, Schepers, & Peres, 2004). Also, the hierarchical IRT model (Glas & van der Linden, 2003) could be considered as belonging to this class, if item categories are considered to define substantive combinations of features. In this section, these models will be reviewed.

The LLTM (Fischer, 1973) belongs to the Rasch family of IRT models, but item difficulty is replaced with a model of item difficulty. In LLTM, items are scored on stimulus features, q_{ik}, which is the score of item *i* on stimulus feature *k* in the cognitive complexity model of items. Estimates from LLTM include η_k, the weight of stimulus feature *k* in item difficulty and θ_j, the ability of person *j*. The probability that the person *j* passes item *i*, $P(X_{ij}=1)$ is given as follows:

$$P(X_{ij}=1 \mid \theta_j, \boldsymbol{q}, \boldsymbol{\eta}) = \frac{\exp(\theta_j - \sum_{k=1}^{K} q_{ik}\,\eta_k)}{1+\exp(\theta_j - \sum_{k=1}^{K} q_{ik}\,\eta_k)} \tag{1}$$

where q_{i1} is unity and η_1 is an intercept. No parameter for item difficulty appears in the LLTM; instead, item difficulty is predicted from a weighted combination of stimulus features that represent the cognitive complexity of the item. That is, the weighted sum of the complexity factors replaces item difficulty in the model. In Janssen *et al's* (2004) model, LLTM-R, a random error term is added to the item composite to estimate variance in item difficulty that is not accounted for by the complexity model.

The 2PL-Constrained model (Embretson, 1999) includes cognitive complexity models for both item difficulty and item discrimination. In this model, q_{ik} and q_{im} are scores of stimulus factors, for item difficulty and item discrimination, respectively, in item *i*. The model parameter η_k is the weight of stimulus factor *k* in the difficulty of item *i,* τ_m is the weight of stimulus factor *m* in the discrimination of item *i* and θ_i is defined as in Equation 1. The 2PL-Constrained model gives the probability that person *j* passes item *i* as follows:

$$P(X_{ij}=1 \mid \theta_j, \boldsymbol{q}, \boldsymbol{\eta}, \tau) = \frac{\exp(\sum_{m=1}^{M} q_{im}\tau_m(\theta_j - \sum_{k=1}^{K} q_{ik}\,\eta_k))}{1+\exp(\sum_{m=1}^{M} q_{im}\tau_m(\theta_j - \sum_{k=1}^{K} q_{ik}\,\eta_k))} \tag{2}$$

where q_{i1} is unity for all items and so τ_1 and η_1 are intercepts. The 2PL-Constrained is the cognitive model version of the 2PL model, since both the item difficulty parameter b_i and the item discrimination parameter a_i are replaced with cognitive models.

Finally, the hierarchical IRT model (Glas & van der Linden, 2003) is similar to the 3PL model, except that the parameters represent a common value for a family of items. The probability is given for person *j* passing item *i* from family *p*, and the item parameters are given for the item family, as follows:

$$P(X_{ij_p}=1 \mid \theta_j, a_{i_p}, b_{i_p}, c_{i_p}) = c_{i_p} + \left(1-c_{i_p}\right)\frac{\exp(a_{ip}(\theta_j - b_{ip}))}{1+\exp(a_{ip}(\theta_j - b_{ip}))} \tag{3}$$

where a_{ip} is item slope or discrimination of item family *p*, b_{ip} is the item difficulty of item family *p*, c_{ip} is lower asymptote of item family *p* and θ_j is ability for person *j*. Thus, in this model, items within family *p* are assumed to have the same underlying sources of item difficulty, but differing surface features. For example, in mathematical word problems, the same essential problem can be presented with different numbers, actors, objects and so forth. Of

course, substituting surface features can create variability within a family and hence the hierarchical model includes assumptions about the distribution of the item parameters.

It should be noted that the original LLTM also can be formulated to represent item families in modeling item difficulty. That is, the q_{ik} represent a set of dummy variables that are scored for items in the same family. The random effects version, LLTM-R, can be formulated with dummy variables and, in this case, would estimate variance due to items within families.

Advantages of LLTM and related models

If LLTM is estimated with item complexity factors that represent theoretically interesting and empirically supported variables, then applying the model to operational tests has several advantages. First, construct validity is explicated at the item level. The relative weights of the underlying sources of cognitive complexity represent what the item measures. Messick (1995) describes this type of item decomposition as supporting the substantive aspect of construct validity. Second, test design for ability tests can be based on cognitive complexity features that have been supported as predicting item psychometric properties. That is, the test blueprint can be based on stimulus features of items that have empirical support. Third, the empirical tryout of items can be more efficiently targeted. Typically, item banks have shortages of certain levels of difficulty. By predicting item properties such as item difficulty, empirical tryout can be restricted to only those items that correspond to the target levels of difficulty. Furthermore, items with construct-irrelevant features can be excluded from tryout if they are represented by variables in the model. Fourth, predictable psychometric properties can reduce the requisite sample size for those items that are included in an empirical tryout (Mislevy, Sheehan & Wingersky, 1993). The predictions set prior distributions for the item parameters, which consequently reduces the need for sample information. Under some circumstances, predicted item parameters function nearly as well as actually calibrated parameters (Bejar *et al*, 2003).

Fifth, a plausible cognitive model provides a basis for producing items algorithmically. Items with different sources of cognitive complexity can be generated by varying item features systematically, based on the cognitive model. Ideally, these features are then embedded in a computer program to generate large numbers of items with predictable psychometric properties (e.g., Embretson, 1999; Hornke & Habon, 1986 ; Adrensay, Sommer, Gittler & Hergovich, 2006). Sixth, a successful series of studies to support the model of item complexity can provide the basis for adaptive item generation. This advantage takes computerized adaptive testing to a new level. That is, rather than selecting the optimally informative item for an examinee, instead the item is generated anew based on its predicted psychometric properties, as demonstrated by Bejar *et al* (2003). Finally, score interpretations can be linked to expectations about an examinee's performance on specific types of items (see Embretson & Reise, 2000, p. 27). Since item psychometric properties and ability are measured on a common scale, expectations that the examinee solves items with particular psychometric properties can be given. However, item estimates based on LLTM and related models go beyond the basic common scale, because the item solving probabilities are related to various sources of cognitive complexity in the items. Stout (2007) views this linkage as extending continuous IRT models to cognitive diagnosis, in the case of certain IRT models.

The cognitive complexity of mathematical problem solving items

Cognitive complexity level and depth of knowledge have become important aspects of standards-based assessment of mathematical achievement. Cognitive complexity level is used to stratify items in blueprints for many state year-end tests and in national tests. However, obtaining reliable and valid categorizations of items on cognitive complexity has proven challenging. For example, although items with greater cognitive complexity should be empirically more difficult, it is not clear that evidence supports this validity claim. Further, rater reliability even on national tests such as NAEP is not reported. Yet another challenge is some systems for defining cognitive complexity (e.g., Webb, 1999) seemingly relegate the predominant item type on achievement tests (i.e., multiple choice items) to only the lowest categories of complexity.

In this study, a system for understanding cognitive complexity of mathematical problem solving items is examined for plausibility. The system is based on a postulated cognitive model of mathematical problem solving which is examined for empirical plausibility using item difficulty modeling. Two approaches to understanding cognitive complexity in items, LLTM and regression modeling, are presented and compared. For both approaches, the development of quantitative indices of item stimulus features that impact processing difficulty is required. Processing difficulty, in turn, impacts psychometric properties. If the model is empirically plausible, support for the substantive aspect of validity (Messick, 1995) or for construct representation validity (Embretson, 1998) is obtained.

Theoretical background

The factors that underlie the difficulty of mathematics test items have been studied by several researchers, but usually the emphasis is to isolate the effects of a few important variables (e.g., Singley & Bennett, 2002; Arendasy & Sommer, 2005; Birenbaum, Tatsuoka, & Gurtvirtz, 1992). The goal in the current study was to examine the plausibility of a model that could explain performance in a broad bank of complex mathematical problem solving items. Mayer, Larkin and Kadane's (1984) theory of mathematical problem solving is sufficiently broad to be applicable to a wide array of mathematical problems. Mayer *et al* (1984) postulated two global stages of processing with two substages each: Problem Representation, which includes Problem Translation and Problem Integration as substages, and Problem Execution, which includes Solution Planning and Solution Execution as substages. In the Problem Representation stage, an examinee converts the problem into equations and then in the Problem Execution stage, the equations are solved. Embretson (2006) extended the model to the multiple choice item format by adding a decision stage to reflect processing differences in the role of distractors (e.g., Embretson and Wetzel, 1987).

Figure 1, an adaptation of an earlier model (Embretson, 2006), presents a flow diagram that represents the postulated order of processes in general accordance with Mayer *et al's* (1984) theory. A major distinction in the model is equation source. If the required equations are given directly in the item, then Problem Execution is the primary source of item difficulty. If the requisite equations are not given directly in the item, then processes are needed to translate, recall or generate equations. Once the equations are available in working mem-

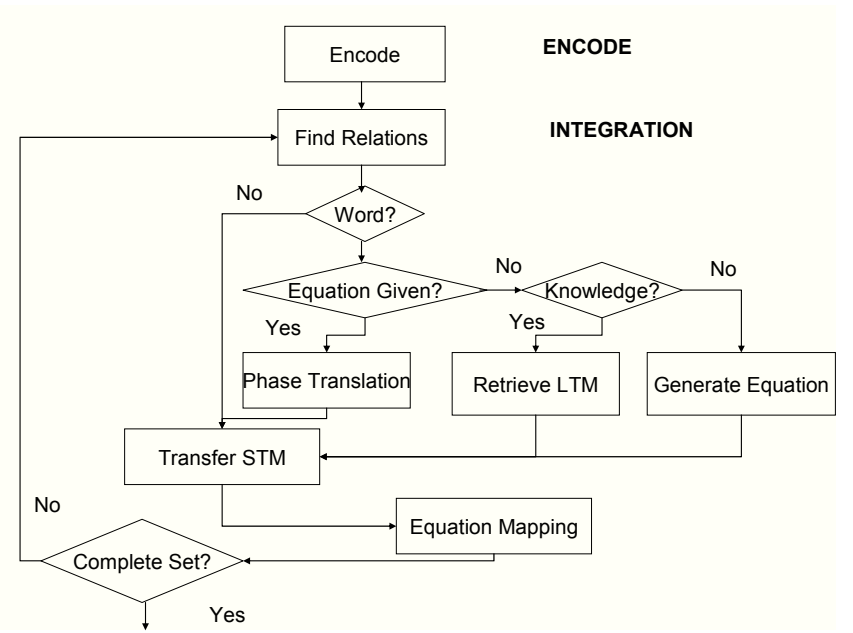

Figure 1:
Cognitive Model for Mathematical Problem Solving Items

ory, the Problem Execution stage can be implemented. Problem Execution involves 1) planning, in which a strategy for isolating the unknowns is developed and implemented, 2) solution execution, in which computations are made to obtain the unknowns, and 3) decision, in which the obtained solution is compared to the response alternatives. Some items do not require solution planning, however, since the equations are given in a format so that only computation algorithms need to be applied.

Several variables were developed in a previous study (Embretson, 2006) to represent the difficulty of the postulated stages in mathematical problem solving. For the Problem Translation stage, a single variable, encoding, was scored as the sum of the number of words, term and operators in the stem. For the Problem Integration stage, several variables were scored to represent processing difficulty for items in which the equation was not given: 1) translating equations from words, 2) number of knowledge principles or equations to be recalled, 3) maximum grade level of knowledge principles to be recalled and 4) generating unique equations or representations for the problem. For some items a special problem representation may be needed; that is, visualization may be required when a diagram is not provided. For the Solution Planning stage, two variables, the number of subgoals required to obtain the final solution and the relative definition of unknowns determine item difficulty. For the Solution Execution stage, the number of computations and the procedural level impact item difficulty. Finally, for the Decision stage, item difficulty is impacted when finding the correct answer involves extensive processing of each distractor. This occurs when the answer obtainable from the stem alone cannot be matched directly to a response alternative.

Method

Test and item bank. A large set of disclosed items were available from the Quantitative section of the GRE. The Quantitative section contains three types of items; Problem Solving, Quantitative Comparison and Data Interpretation. A Problem Solving item consists of a stem that defines the problem and five unique response choices. The stems range in length and type; some stems are highly elaborated word problems while others contain equations or expressions. A Quantitative Comparison item consists of a short stem and two columns, A and B, that contain either numbers or equations. Each item has the same four response alternatives; "The quantity in A is greater", "The quantity in B is greater", "The quantities are equal", and "The relationship cannot be determined". Finally, Data Interpretation items consist of a graph, table or chart, a short stem that poses a question about the display and five unique response alternatives.

For each item, the GRE item bank parameters were available. In the current study, only the Problem Solving items were modeled because they are used widely on many high stakes tests to measure achievement or ability.

Design. Eight test forms had been administered in a previous study to collect item response time data for developing the cognitive model for the mathematical problem solving items (Embretson, 2006). However, item response data were also available and had not been previously analyzed. Each test form contained 43 items, of which 12 items were linking items and the remaining items were a mixture of Problem Solving items and Quantitative Comparison items. A total of 112 Problem Solving items were included across the eight forms.

Participants. The participants were 534 undergraduates from a large Midwestern University who were enrolled in an introductory psychology course. The participants were earning credits as part of a course requirement.

Procedures. Each participant was randomly assigned a test form. All test forms were administered by computer in a small proctored laboratory. The test administration was not speeded as participants were allowed up to one hour to complete the test form. Nearly all participants completed the test in the allotted time.

Cognitive complexity scores. The items were scored for cognitive complexity by multiple raters. All items were outlined for structure prior to scoring. The variables were scored as follows: 1) Encoding, a simple count of the number of words, terms and operators in the item stem, 2) Equation Needed, a binary variable scored "1" if the required equation was not included in the item stem, 3) Translate Equations, a binary variable scored "1" if the equation was given in words in the item stem, 4) Generate Equations, a binary variable scored "1" if the examinee had to generate a unique representation of the problem conditions, 5) Visualization, a binary variable scored "1" if the problem conditions could be represented in a diagram that was not included, 6) Maximum Knowledge, the grade level of the knowledge required to solve the problem (scored from National Standards), 7) Equation Recall Count, the number of equations that had to be recalled from memory to solve the problem, 8) Subgoals Count, the number of subgoals that had to be solved prior to solving the whole problem, 9) Relative Definition, a binary variable scored "1" if the unknowns were defined relative to each other, 10) Procedural Level, the grade level of the required computational procedures to solve the problem, 11) Computational Count, the number of computations required to solve the problem and 12) Decision Processing, scored "1" if extended processing of the distractors was required to reject all but the correct answer.

Results

Descriptive statistics. The Rasch model parameters were calibrated using the twelve common items to link item difficulties (b_i) across forms using BILOG-MG. Estimates were scaled by fixing item parameters (Mn = 0, Slope = 1), according to typical Rasch model procedures. An inspection of the goodness of fit statistics for items, based comparing expected versus observed response frequencies, indicated that only two of the 112 items failed to fit the Rasch model (p's < .01). Lowering the criterion for misfit resulted in only one more additional item that failed to fit (p <.05). Thus, the Rasch model was a good fit to the item response data.

Since the sample differed from the target GRE population, it was desirable to assure that the difficulties were appropriate for the sample. Thus, items were selected for appropriate item difficulties within a specified range of estimated item difficulties ($-1.80 < b_i < 1.80$). Eleven items of the 112 were eliminated using this criterion, thus leaving 101 items for analysis. The classical item difficulty statistic, *p-value*, for the remaining items then fell within an acceptable range ($.10 < p_i < .90$).

Table 1 shows the mean item difficulty for the estimates from the sample and from the item bank parameters. The means and standard deviations differ, but these differences are due to differences in the population of examinees, the standardization procedures (i.e., scaling the solution to the items or to the population ability distribution) and the model that was estimated (Rasch versus 3PL). Nonetheless, the item difficulty estimates from the sample were highly correlated with the item bank estimates (r = .834), thus supporting general comparability of the task across the two populations.

Table 1 also presents the means for the cognitive model variables. It can be seen that Encoding is fairly complex across items, with an average of almost 29 words, terms and operators per item stem. Equation Needed, Translate Equations, Generate Equations and Visualization are all binary variables and can be interpreted as proportions. Hence, a high proportion of items require the examinee to produce an equation, many problems require either translating or generating equations, while few items require visualization. Maximum Knowledge indicates that most items required 7^{th} grade knowledge or less, while Equation Recall Count indicates that the mean number of equations to be recalled was somewhat less than one. The mean for Subgoals Count was also somewhat less than one, which indicates that many problems have subgoals. The mean for Relative Definition, a binary variable, indicated that many problems have relatively defined unknowns and consequently require procedures for solving simultaneous equations. The mean for Procedural Level, a contrast coded variable, was consistent with the average item involving no higher level than computations with fractions. The mean for Computation Count indicated that most problems involved moderate numbers of computations. Finally, the mean for Decision Processing, a binary variable, indicated that only a small proportion of items involved extensive comparisons of the distractors.

Table 2 presents the correlations of item difficulty with the cognitive model variables. Although the two estimates of item difficulty were obtained from both different populations and different IRT models (i.e., Rasch model for the undergraduate sample and the 3PL model for the item bank calibrations, it can be seen that the correlations are quite similar with the cognitive model variables. Significant positive correlations with item difficulty were obtained for most variables. That is, Encoding, Equation Needed, Translate Equations, Gen-

erate Equations, Visualization, Maximum Knowledge, Equation Recall Count, Relative Definition and Decision Processing had significant positive correlations with both estimates of item difficulty (p's < .05). Subgoal Count had a significant positive correlation with the

Table 1:
Descriptive Statistics on Model Variables

	Mean	Std. Deviation	N
Item Difficulty from Item Bank	.31	.94	101
Item Difficulty Sample Estimate	-.00	.79	101
Encoding	28.63	13.67	101
Equation Needed	.89	.31	101
Translate Equations	.38	.49	101
Generate Equatiions	.33	.47	101
Visualization	.08	.27	101
Maximum Knowledge	6.98	1.45	101
Equation Recall Count	.75	1.03	101
Subgoal Count	.76	1.00	101
Relative Definition	.44	.50	101
Procedural Level	-1.12	1.75	101
Computation Count	3.74	2.39	101
Decision Processing	.06	.24	101

Table 2:
Correlations of Model Variables with Item Difficulty Estimates

	Item Difficulty from Item Bank	Item Difficulty Sample Estimate
Item Difficulty Item Bank	1.000	.834**
Item Difficulty Sample	.834**	1.000
Encoding	.355**	.328**
Equation Needed	.305**	.232**
Translate Equations	.247**	.302**
Generate Equations	.479**	.435**
Visualization	.175*	.257**
Maximum Knowledge	.201*	.205*
Equation Recall Count	.144+	.166*
Subgoal Count	.165*	.141+
Relative Definition	.348**	.323**
Procedural Level	-.041	.028
Computation Count	.085	.091
Decision Processing	.362**	.343**

+ p < .10, * p < .05, ** p < .01

item bank estimates of item difficulty and a marginally significant positive correlation with the sample estimates of item difficulty. The highest single correlation with item difficulty for both estimates was for Generate Equations. Procedural Level and Computation Count did not correlate significantly with either estimate of item difficulty.

Regression modeling of cognitive complexity. Each estimate of item difficulty was modeled using hierarchical regression. Table 3 presents the hierarchical regression models by stage for each estimate of item difficulty. The relevant cognitive variables for each stage were entered as a block in a hierarchical regression, with the order of entry into the model corresponding to the temporal order of the stages. It can be seen that the results are quite similar with one minor exception. That is, for both estimates of item difficulty, the Encoding, Integration, Solution Planning and Decision stages had significant contributions to prediction (p's < .05). For the Solution Execution stage, however, a marginally significant effect (p = .076) was obtained for item difficulty estimates from the item bank, but no significant effect for the sample estimates of item difficulty.

The patterns of significance for the regression coefficients of the individual cognitive variables within the stages, however, did vary across the two item difficulty estimates, even though the scored model variables were identical. For the item difficulty estimates from the item bank, shown on Table 4, significant regression coefficients were found for Encoding, Generate Equation and Decision Processing (p's < .05). Marginally significant regression

Table 3:
Hierarchical Regression Modeling by Postulated Processing Stage

Model for Item Difficulty Estimates from Item Bank

			Change Statistics				
Model	R	R Square	R Square Change	F Change	df1	df2	Sig. F Change
1 Encoding	.355	.126	.126	14.258	1	99	.000
2 Integration	.596	.356	.230	5.529	6	93	.000
3 Solution Planning	.625	.391	.036	2.654	2	91	.076
4 Solution Execution	.633	.401	.009	.703	2	89	.498
5 Decision	.672	.451	.051	8.115	1	88	.005

Model for Sample Estimates of Item Difficulty

			Change Statistics				
Model	R	R Square	R Square Change	F Change	df1	df 2	Sig. F Change
1 Encoding	.328	.107	.107	11.919	1	99	.001
2 Integration	.591	.349	.242	6.975	5	94	.000
3 Solution Planning	.604	.365	.016	.759	3	91	.520
4 Solution Execution	.617	.381	.016	1.166	2	89	.316
5 Decision	.660	.435	.054	8.441	1	88	.005

Table 4:
Regression Coefficients for Modeling of Item Difficulty Parameters from the Item Bank

Item Predictor	b Coeff.	SE_b	t	Prob.
Constant	-.678	.551	-1.229	.222
Encoding	.015	.007	2.227	.028
Equation Needed	.195	.268	.729	.468
Translate Equations	.070	.197	.354	.724
Generate Equations	.663	.184	3.613	.001
Visualization	.533	.281	1.899	.061
Maximum Knowledge	-.008	.080	-.094	.925
Equation Recall Count	-.065	.124	-.521	.604
Subgoals Count	.203	.121	1.672	.098
Relative Definition	.319	.188	1.699	.093
Procedural Level	.039	.046	.836	.405
Computation Count	-.024	.039	-.622	.535
Decision Processing	.999	.353	2.826	.006

coefficients were found for Visualization, Subgoals Count and Relative Definition. Thus, some support was found for four of the model variables and marginal support was found for two more variables.

For the item difficulty estimates from the sample, presented on Table 5, significant regression coefficients were observed for Generate Equations and Decision Processing and marginally significant coefficients were observed for Visualization and Encoding. Thus support was found for only two of the model variables and marginal support was found for another two variables.

Table 5:
Regression Coefficients for Cognitive Model on Undergraduate Sample

Item Predictor	b Coeff.	SE_b	t	Prob.
Constant	-.513	.472	-1.086	.280
Encoding	.010	.006	1.701	.093
Equation Needed	-.050	.229	-.217	.829
Translate Equations	.291	.170	1.712	.090
Generate Equations	.532	.159	3.337	.001
Visualization	.584	.254	2.295	.024
Maximum Knowledge	-.025	.068	-.367	.715
Equation Recall Count	.040	.107	.375	.709
Subgoal Count	.086	.104	.825	.412
Relative Definition	.169	.162	1.043	.300
Procedural Level	.057	.040	1.434	.155
Computation Count	-.009	.034	-.280	.780
Decision Processing	.878	.302	2.905	.005

LLTM analysis of cognitive complexity. Maximum likelihood estimates of LLTM parameters from the raw item data were based on formulating the model as a non-linear mixed model (DeBoeck & Wilson, 2004). The person parameters were specified as random variables from a standard normal distribution, ~N(0,1). The item model variables were specified as fixed effects. All models were estimated with adaptive Gaussian quadrature to obtain marginal maximum likelihood estimates for the parameters. Three models were specified as alternative LLTMs: 1) the Rasch model, with 101 dummy variables and no intercept, 2) the cognitive model with 12 estimates plus an intercept and 3) a null model, with a single estimate of item difficulty which was used for comparison purposes. Since the estimates were scaled to the person distribution, a constant slope (*a*) was also estimated for each model. LLTM estimated with these constraints yields the same likelihood as LLTM parameters estimated with *a* fixed to 1.0 and a freely estimated variance. The item parameters may be rescaled by *a*. Linking across forms was obtained through the common items on each form.

Since the Rasch model had been previously established as appropriate for the data (see above), the main concern was the relative fit of the three models. As expected, the Rasch model was the best fitting model, as it had the smallest value on the Akaike Information Criterion *(-2lnL*= 10,859, *AIC* = 11,063). The LLTM, with 12 estimates for the model variables and an intercept, was the second best model *(-2lnL*=11,296, *AIC* = 11,324). Finally, the null model yielded the worst fit *(-2lnL*=11,778, *AIC* = 11,782).

The relationship of the LLTM predictions of item difficulty to the Rasch model item difficulties was examined in two ways. First, a fit index based on the log likelihoods of the alternative models, the delta fit index (Embretson, 1995) was computed. The delta index ranges from 0 to 1 and is based on the comparison of the relative likelihood of the null model to the likelihood of the LLTM and the Rasch model. Moderately strong fit was obtained ($\Delta^{1/2}$=.724), which is comparable in magnitude to a multiple correlation. Second, a scatterplot of the Rasch item difficulties versus the LLTM predictions was prepared, as shown on Figure 2. The Rasch estimates are shown as error bars based on a 90 percent confidence interval defined by the associated standard errors for each estimate. The LLTM prediction for each item is shown as a dark circle. It can be seen that for a majority of the items, the LLTM predictions fall within the 90 percent confidence interval for the Rasch item difficulties.

Table 6 presents the estimates of the cognitive model coefficients that were from the LLTM. Eight cognitive model variables, Encoding, Equation Needed, Translate Equation, Generate Equations, Visualization, Subgoals Count, Relative Definition and Decision Processing, are statistically significant (p's < .05). An additional cognitive variable, Procedural Level, has borderline statistical significance (p=.0516). Only three model variables, Maximum Knowledge, Equation Recall and Computation Count, failed to reach statistical significance.

Both direct estimates of the coefficients for the model variables (η) and their associated standard errors are shown on Table 6. It can be seen that the standard errors for the model coefficients are much smaller for the LLTM than for the regression modeling coefficients shown on Table 4 and Table 5. In LLTM, the size of the standard errors is related to the number of persons in the dataset. For the regression modeling, the standard error depends on the number of items since item statistics are the target of the modeling. Table 6 also presents rescaled estimates (η*) of the model coefficients obtained from LLTM. The rescaled estimates (η*) may be compared to the regression coefficients on Table 5. Rescaling was necessary due to the differing manner of model identification in the two analyses, as described

above. Table 6 also shows that the rescaled estimates for most model variables are similar in pattern and general magnitude to the regression coefficients shown on Table 5. However, the exact estimates do vary, as would be expected from the bias produced by modeling statistics rather than raw data.

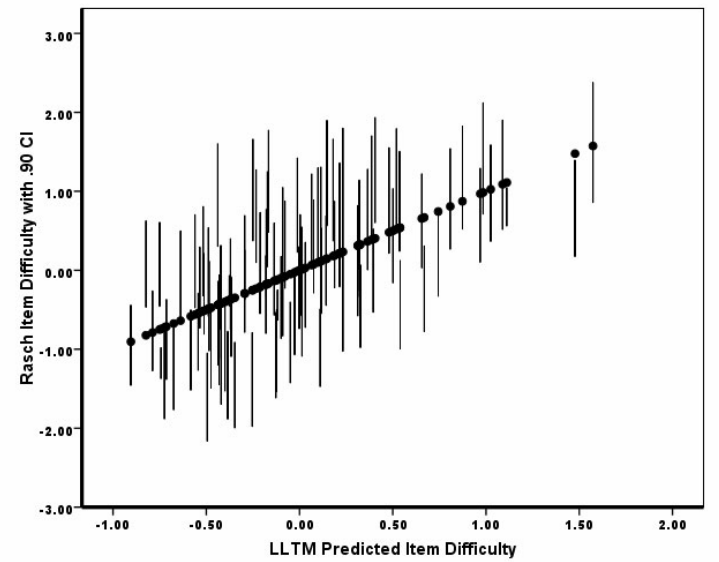

Figure 2:
Prediction of Rasch Item Difficulty from LLTM

Table 6:
Parameter Estimates for the LLTM

Item Predictor	η	se	t	Prob.	η *
Constant	.116	.208	.56	.5769	-.549
Discrimination Constant	.864	.041	21.09	<.0001	1.000
Encoding	.015	.003	5.51	<.0001	.013
Equation Needed	-.355	.095	-3.72	.0002	-.307
Translate Equation	.291	.083	3.51	.0005	.251
Generate New Equation	.472	.072	6.53	<.0001	.408
Visualization	.616	.128	4.83	<.0001	.532
Maximum Knowledge	-.023	.029	-.78	.4371	-.020
Equation Recall Count	.015	.047	.03	.7521	.013
Subgoals Count	.135	.049	2.76	.0060	.117
Relative Definition	.285	.076	3.73	.0002	.246
Procedural Level	.037	.019	1.95	.0516	.032
Computation Count	-.004	.017	-.23	.8154	-.003
Decision Processing	1.23	.186	6.63	<.0001	1.063

Item design. As noted above, a plausible model of item complexity using LLTM permits items to be banked by their sources and levels of complexity. In the cognitive model for mathematical problem solving that was developed, the processes can be regarded as compromising two global stages, Problem Representation and Problem Execution. The sources of item complexity can be obtained for the global stages by combining the variables according to the LLTM weights that are included in the sub-stages. Thus, the predicted problem complexity due to Problem Representation, b'_{PR}, can be obtained as follows:

b'_{PR} = .015(Encoding) + (-.355)(Equation Needed) + .291(Translate Equation) + .472(Generate Equation) + .616(Visualization) + (-.023)(Maximum Knowledge) +.015(Equation Recall).

Similarly, the predicted problem complexity due to Problem Execution difficulty, b'_{PD}, can be obtained as follows:

b'_{PD} = .135(Subgoals Count) + .285(Relative Definition) + .037(Procedural Level) + (-.004)(Computation Count) + 1.23(Decision Processing).

Applying these weights to the scored variables, q_{ik}, resulted in scores for each item on the global stages. The means for Problem Representation (Mn = .277, SD = .401) and Problem Execution (Mn = .241, SD = .349) are very similar in magnitude. The correlation between the sources of item complexity was statistically significant, but relatively small in magnitude ($r = .255$, $p<.01$). Figure 2 presents a scatterplot of the two predicted sources of item difficulty. The means for each variable are shown as reference lines. It can be seen that items fall

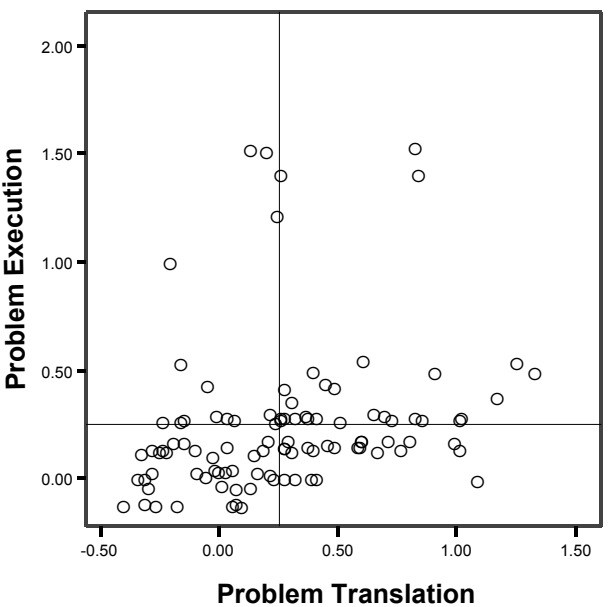

Figure 3:
Scatterplot of Two Major Sources of Item Complexity

within each section of the scatterplot. Thus, there are items which involve primarily Problem Representation or Problem Translation as sources of difficulty, while other items are more balanced for sources of difficulty.

Discussion

The results have several implications. First, the results support the validity of the cognitive model to understand the substantive aspect of validity (Messick, 1995) for items from a major test of quantitative ability. The postulated model of cognitive complexity for mathematical problem solving was supported from the results of both the regression approach and the LLTM approach. The cognitive model accounted for about half the variance of item difficulty. This is a strong level of prediction for modeling item difficulty in a broad bank of existing mathematical problem solving items. The items in the current study were quite diverse in content, syntax and form, as well as in mathematical requirements. Higher levels of prediction (Enright, Morley & Sheehan, 1999; Bejar et al, 2003; Embretson & Daniel, 2008) have been reported for items that are created by systematically varying features of existing items (i.e., item models). That is, the items that are created are identical except for those features that are designed to vary. The predictions obtained from these models, consequently, extend only to the difficulty of items produced from the item models which were varied in the study, not to the difficulty of items in broader item bank.

Second, the results support the potential of the cognitive model as item design principles for cognitive complexity. That is, support was found for the variables that were hypothesized to impact complexity in the postulated processing stages. However, the LLTM approach led to a more consistent and powerful estimation of the impact of the complexity variables in the cognitive model. Most of the model variables were significant predictors in LLTM, which suggests that they can be varied to produce items with different sources and levels of complexity. In contrast, the hierarchical regression modeling approach led to support for most of the postulated processing stages, but was unable to untangle the contribution of specific variables. Only two item complexity variables, generating mathematical representations and extensive processing of the distractors, were indicated as consistently significant predictors of item difficulty from both the sample estimates and the item bank parameters. Since LLTM makes full use of the data in estimating the standard errors of the parameters, it led to more powerful tests of the model variables. Results from a recent study supported the validity of the item complexity variables that were identified with the more powerful LLTM analysis. That is, Embretson and Daniel (2008) found that controlled manipulations of several variables in the cognitive model impacted item difficulty.

Third, the results support the potential of selecting mathematical problem solving items for specific levels and sources of cognitive complexity. The coefficients for LLTM are useful for item banking by sources of cognitive complexity, particularly since they are item response theory model estimates. To show the potential for test design, item difficulties were estimated for the two global stages of information processing in the cognitive model, which represents a composite of the corresponding cognitive model variables. Interesting, Problem Representation, which includes encoding the item and producing an equation, was approximately equal in difficulty in the items as Problem Execution, which includes planning to isolate unknowns and computing the results. However, although these two stages were generally equally difficult in

the item set as a whole, substantial variability was found between items. The scatterplot showed that items could be selected to represent about an equal balance of the sources of difficulty, if desired. However, items also could be selected to represent primary Problem Representation difficulty or Problem Execution difficulty. This is an important design issue. One would expect ability test items to consistently involve difficulty in the Problem Representation phase, which probably involves greater levels of reasoning. In contrast, mathematical achievement tests may emphasize Problem Execution more predominantly.

Fourth, the level of prediction obtained may also be sufficient to select items for operational use without further tryout. Simulation studies have shown that although using predicted item properties in place of calibrated properties increases the standard errors of score estimates (see Mislevy, Sheehan and Wingersky, 1985; Embretson, 1999; Bejar, 2003), the increased error may be easily compensated by a slightly longer test. Further research is clearly needed on this issue.

Finally, it should be noted that LLTM, rather than the regression modeling approach, should be applied whenever possible. The main advantage is that the standard errors will be appropriate so that the effects of the cognitive model variables may be untangled. Unfortunately, applying LLTM is not always feasible in practical testing situations, due to the unavailability of the raw data. It is important that the researcher be aware of the disadvantages in interpreting the results.

In summary, the application of LLTM to mathematical problem solving items from a widely used test of quantitative ability clearly supported a processing model of item complexity. The results have several implications for item and test design, as well as for construct validity.

References

Arendasy, M., Sommer, M., & Ponocny, I. (2005). Psychometric approaches help resolve competing cognitive models: When less is more than it seems. *Cognition & Instruction, 23*, 503–521.

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3).

Birenbaum, M., Tatsuoka, K. K., & Gurtvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16(4)*, 353 - 363.

DeBoeck, P., & Wilson, M.(2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

Embretson, S. E. (1996). Cognitive design systems and the successful performer: A study on spatial ability. *Journal of Educational Measurement, 33*, 29-39.

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

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*, 407-433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and Practice*. Mahwah, NJ: Erlbaum.

Embretson, S. E. (2006). *Cognitive models for the psychometric properties of GRE quantitative items.* Final Report. Educational Testing Service.

Embretson, S. E., & Daniel, R. C. (2008). *Designing cognitive complexity in mathematical problem solving items.* Paper presented at the annual meeting of the American Educational Research Association. New York, NY: March.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. Invited article for *Journal of Educational Measurement, 38*, 343-368.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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

Enright, M. K., Morley, M., & Sheehan, K. M. (1999). Items by Design: The Impact of Systematic Feature Variation on Item Statistical Characteristics. *Applied Measurement in Education, 15 (1),* 49-74.

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

Freedle, R., & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (ETS Research Report RR91– 29). Princeton, NJ: ETS.

Freedle, R., & Kostin, I. (1993).The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*, 133 – 170.

Glas, C. A. W., & Van der Linden, W. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247-261.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42,* 351–373.

Gorin, J., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30,* 394-411.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*, 369-380.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York: Springer.

Mayer, R. E., Larkin, J., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem solving ability. In R. Sternberg's (Ed.), *Advances in the psychology of human intelligence*, *V2* (pp. 231-273).

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-76.

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning, 12*(1), 62-90.

Singley, M., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds), *Item Generation for Test Development* (pp. 361-384). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Sheehan, K. M., & Ginther, A. (2001). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section.* Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education.

Spada, H., & McGaw, B. The assessment of learning effects with linear logistic test models. In Embretson, S. E. (Ed.). *Test design: Developments in psychology and psychometrics* (pp. 169-193). New York: Academic Press.

Stout, W. (2007). Skill diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*, 313-324.

Webb, N. L. (1999). *Criteria for alignment of expectations and assessments in mathematics and science education.* Research Monograph No. 6. Wisconsin Center for Educational Research. Madison, WI.