Running head: MODELING LANGUAGE COMPONENTS IN MATH ITEMS

Modeling Language Components in Mathematical Items Using Multiple-Random-Effects IRT Models

Xiangdong Yang

Indiana University, Bloomington

Pui Chi Chiu

University of Kansas

Please address correspondence to Xiangdong Yang, Dept. of Counseling and Educational Psychology, Indiana University, 201 N. Rose Avenue, Bloomington, IN 47405; e-mail: yang50@indiana.edu.

Abstract

Language components in mathematical word problems might have some profound impacts on the problem solving processes engaged by examinees. Such impacts usually do not occur uniformly across examinees, but rather are the results of the interactions between the characteristics of the mathematical item and the cognitive properties of the particular examinee. This study examined such impacts through a multiple random-effects IRT modeling approach, which has the benefit of modeling the individual-specific impacts of the language components on mathematical problem solving, in addition to the average effects of such components across the testing sample. Results showed that substantial variations among examinees of the impacts from such language components were found for the examinee sample.

Keywords: language component, mathematical word problems, random weight LLTM, cross random-effects LLTM

Modeling Language Components in Mathematical Items

Using Multiple-Random-Effects IRT Models

Comprehending and solving mathematical word problem is generally perceived as a complex cognitive activity that involves several processes (Mayer, 1984; Schoenfeld, 1983). From an information-processing perspective, Schoenfeld formulated a processing model of mathematical problems solving that includes five sub-processes: reading, analysis, exploration, planning/implementation and verification. Along the same vein of study, Mayer identified four general cognitive processes of mathematical word problem solving: problem translation, problem integration, solution planning and solution execution. Problem translation is to translate the problem statements into mathematical propositions, such as equations, whereas problem integration is to put various elements of the problem statements, questions, and explicit or implicit constraints into a coherent representation of the problem structure. After an integrated representation was formed, a mathematical problem can be solved through developing and monitoring a solution plan and executing such a plan. When problem is presented with language statements, translation of such language statements into mathematical propositions must be carried out correctly before relevant information can be integrated into a coherent representation of the problem. It is at the stage of problem translation that the complexity of language statements plays a profound role.

This study aims to investigate the effects of language components in mathematical items on examinees' performances and applies several item response theory (IRT) models with multiple random effects. This approach starts from searching for a cognitive processing model for mathematical problem solving, and then identifies specific item stimulus features in the problem statements to represent the language components according to the model. By employing

IRT models with multiple random effects, this approach allows researcher to model the individual-specific impacts of the language components on mathematical problem solving, in addition to the average effects of such components across the testing sample. In the following sections, we first reviewed research that address different aspects of language components in mathematical word problems, and then discussed different approaches that have been taken to model impacts of language components on mathematical problem solving as well as their limitations. The modeling approach in the current study was described next and was followed by an empirical study of the impacts of language components on mathematical problem solving in a 3rd grade mathematical test. Results from the study were then described and discussed, along with some general discussion of the modeling approach in this study.

Language Components in Mathematical Word Problems

Research has shown that the processes of understanding language statements in mathematical problems are compatible with general principles of text comprehension (Kintsch & Greeno, 1985). Based on their cognitive processing model of reading comprehension, Embretson and Wetzel (1987) described test comprehension process as consisting of two major sub-processes: lexical encoding and coherence processes. Lexical encoding refers to the process of converting the visual text stimuli into a meaningful representation, whereas coherence processes refers to the process of integrating various pieces of representations from text and/or prior knowledge into a coherence semantic network. The difficulty of lexical encoding is mainly affected by the surface structure of the text such as number of words, word familiarity, etc, whereas the coherence processes are strongly affected by the propositional density of the text, which is the number of propositions divided by the number of words.

Many research have been done to modify the surface structures of language statement in mathematical test items (Abedi & Lord, 2001; Hanson, Hayes, Schriver, leMahieu, & Brown, 1998; Kopriva, 1999; Johnson & Monroe, 2004). Research has shown significant impacts on item difficulty of certain surface features such as ambiguous words, complex verbs and math vocabulary (Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Although modifications of such surface features improve performance of examinees with low language proficiency, such improvements are generally small in magnitudes (Abedi & Lord, 2001; Johnson & Monroe, 2004).

According to Kintsch and van Dijk (1978), a proposition consists of concepts, which include one predicate and one or more arguments. An argument can be either a concept or an embedded proposition, which is either the subject or object in the proposition. A predicate consists of a verb, a modifier or a connective. For example, the sentence “Both John and Mary liked their Math teacher” can be analyzed into three propositions (see Table 1).

<Insert Table 1 about here>

Embretson and Wetzel distinguished three types of propositions: modifier proposition, predicate proposition, and connective proposition. For example, proposition 1, 2 and 3 in Table 1 is a modifier, a connective, and a predicate proposition, respectively. While in general propositional density is related to text comprehension difficulty, densities of different types of propositions showed substantively different effects. For example, both connective propositional density and modifier propositional density show significant impacts on difficulty of text comprehension, whereas predicate propositional density does not.

Similar findings about propositions were also found from research on mathematical problem solving. Mayer (1984) differentiated three types of propositions in mathematics word

problem: assignment, relation and question. An assignment proposition assigns a value to a variable, such as “A book costs $65”. A relation proposition expresses a quantitative relation between two variables, such as “John has 5 more apples than Mary”. A question proposition raises a question about the value of a variable, such as “how much is the total?” According to Mayer, the relational proposition imposes the biggest difficulty among the three for examinees to comprehend.

Modeling Impacts of Language Components in Mathematical Items

Several approaches have been reported in the literature to investigate the impact of language components in mathematical problems (Aiken, 1972; Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985; Shaftel, et al, 2006; Wu & Adams, 2006). Early research addressed this issue through studying the relationship between reading abilities of examinees and their performances on mathematical word problem solving (see Aiken, 1972 for a comprehensive review). Under the hypothesis that reading ability affects mathematical problem solving, early studies commonly focused on obtaining measures of general reading ability and mathematical ability for children in the intermediate grades and simply correlated them. Although the correlation coefficients were generally positive (range from .40 and .86), such findings were open to various interpretations, however. One of the major rival interpretations was that the positive correlation between examinees’ reading abilities and mathematical abilities was due to the fact that both variables were associated with their general intelligence. Some researchers suggested changing the measures of general reading ability to those of specific reading abilities such as vocabulary, comprehension, syntax and spelling (Johnson, 1949). But it did not resolve the issue because (a) different measures of specific reading abilities yielded inconsistent predictive validity on examinees’ performances of mathematical problem solving

and (b) the general methodology of first-order correlation or partial correlation between math ability and reading ability did not provide a rigorous approach to understand what specific aspects of language components impact on examinees' mathematical problem solving.

Instead of focusing on the correlational patterns of reading ability and mathematical ability among examinees, an alternative approach is to study the internal structure of the mathematical test items through factor analysis (DeGuire, 1985; Gorsuch, 1983; Wu & Adams, 2006). Theoretically, if substantial amount of individual differences exist in the testing sample on the language components in mathematical items, a language factor would be identified in the resulting factorial structure of the mathematical test. If, on the other hand, all members in the sample can successfully perform the problem translation phase of the mathematical problem solving, there would be no such a language factor in the resulting factorial structure. For this reason, factor analysis, either exploratory or confirmatory, provides an approach to investigate the individual differences on the impacts of language components in mathematical items. However, factors identified from a factor analysis of empirical data are hard to interpret. Whether an identified factor reflects impact of language components in mathematical tests or not is open to debate. For the same dataset, different factoring or rotation methods in factor analysis might lead to different results and consequently different interpretations. Even when the same interpretation can be achieved, a language factor does not provide much information about what specific aspects of such language components that produce the individual differences and how they are related to the cognitive processes of mathematical problem solving (Snow & Lohman, 1989).

Another way to study impacts of language components in mathematical items is (a) to identify specific features of language statements in those items though task analysis and then (b)

to investigate the effects of such features on item difficulties and, consequently, student performances (Abedi & Lord, 2001; Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985; Shaftel, et al, 2006). There are two different approaches within this line of research. One approach is to focus on the surface linguistic features of the language statements in mathematical items (Abedi & Lord, 2001; Shaftel, et al, 2006). By modifying nonmath vocabularies and linguistic structures in the items such as word frequency, voice of verb phrases, etc, researchers in this approach hope to simplify the requirement of language comprehension while maintaining the underlying mathematical structure of the item. Since modifications made in such ways may not necessarily link to the cognitive processes of mathematical problem solving, linguistically simplified items may or may not achieve what the researchers hope for. Although in general linguistically simplified items become easier than before, there are situations that linguistically simplified items are even more difficult (see Shaftel, et al, 2006 for a detailed survey of the relevant literature).

In contrast, the second approach within this line of task analysis relies heavily on cognitive processing models of mathematical problem solving (Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985). For example, based on Mayer's (1984) cognitive theory of mathematical problem solving, Embretson (2004) analyzed the psychometric properties of GRE quantitative items with the aim to identify and understand the levels and sources of cognitive complexity involved in those items. In applying this approach, Embretson represented GRE mathematical problem solving in five distinct stages: encoding, integration, planning, solution execution and decision. Different sets of item characteristics were identified to represent cognitive processes within each of the stages and items were then coded in terms of the set of item characteristics. Using linear regression model or linear latent trait test model (LLTM,

Fischer, 1973), the impacts of different item characteristics, including linguistic features, on the difficulties of the items can be modeled and calibrated. One important advantage of the model-based task analysis approach is that it is construct-oriented. Through linking item features with cognitive processes of item solution and empirically calibrating their impacts on item difficulty, the model-based task analysis approach provides a viable method of investigating the underlying construct representation in the mathematical tests (Embretson, 1983).

Although the model-based task analysis approach has a more principled theoretical basis, it assumes that the impacts of identified item features on item difficulty are constant across examinees. In other words, the effects of such item variables are treated as fixed across examinees, which imply that all examinees engage the same problem solving strategy. Such an assumption may not be held in some cases. For example, Sebrechts and his associates (1996) investigated fixed effects of item features on difficulties of GRE algebra word problems as well as the relation between problem statements and examinees' problem solving strategies or errors. Results from their study suggested that, instead of engaging the same problem solving strategy, examinees with different proficiency levels tend to use different strategies. Moreover, initiation and engagement of a specific strategy to solve an item appeared to depend partially on the characteristics of the particular mathematical item, including the surface context of the item and its linguistic structure. The occurrences of errors that lead to incorrect answers are generally associated with the related strategy employed and, consequently, differ among examinees with different proficiency levels. These findings suggested that impacts of item features did not occur uniformly across examinees, but rather varied based on the interactions between the characteristics of the mathematical item and the cognitive properties of the particular examinee.

Therefore, alternative approach needs to be formulated in order to model individual-specific impacts of item features on item difficulties.

The Modeling Approach for the Current Study

Modeling language components in mathematical items in the current study was based on the assumption that, unless all examinees in the sample succeeded in the problem translation phase, impacts of such language components would vary across examinees. In other words, impacts of language components in mathematical items were treated as random across examinees. Therefore, under this assumption, it is possible to estimate an individual-specific parameter for impacts of language components on mathematical item difficulty.

The modeling approach in the current study will follow the model-based task analysis (in this case, item analysis) approach to identify specific language components in a mathematical item. Specifically, the current approach starts searching for a cognitive processing model for mathematical problem solving. Within the model, phases or processes that are linked to language components are identified. For example, the cognitive theory of mathematical problem solving developed by Mayer (1984) has been widely recognized and extensively studied (Barnett & ceci, 2002; Embretson, 1995; Hall, Kibler, & Wenger, 1989; Jonassen, 2003; Lucangeli, Tressoldi, & Cendron, 2002). As mentioned earlier, this theory decomposes mathematical word problem solving into four cognitive processes: problem translation, problem integration, solution planning and solution execution. According to Mayer, problem translation phase requires an examinee to have some knowledge of language and facts. Because both problem solving phases and language components identified at this stage are theoretical concepts, they may not directly link to the specific item features in the problem statements. Therefore, there is a need to identify specific item features in the problem statements to represent the language components. Empirical studies

should be surveyed as to what those features are and how they affect the process of translating the problem statements into mathematical representation.

Based on the set of item features identified to represent language components in mathematical items, IRT models with multiple random effects can be applied to model their individual-specific impacts on item difficulty. Before introducing IRT models with multiple random effects, it is beneficial to first discuss one-parameter logistic model (Rasch model; Rasch, 1960) and LLTM within a generalized linear mixed model (GLMM) framework (McCulloch & Searle, 2001). It is now widely recognized that IRT models can be formulated as GLMMs or nonlinear mixed models (NLMM) (Boeck & Wilson, 2004; Rijmen, Tuerlinckx, Boeck, & Kuppens, 2003). Within the GLMM framework, the Rasch model can be given as follows:

$$logit\left[P\left(X_{ij}=1|\,\theta_i,b_j\right)\right]=\theta_i-b_j\,, \qquad (1)$$

where $P\left(X_{ij}=1|\,\theta_i,b_j\right)$ indicates the probability of examinee $i\,(i=1,2,...I)$ responds correctly to a dichotomous item $j\,(j=1,2,...J)$; and $X_{ij}=1$ for correct response and 0 otherwise; and θ_i is the person parameter of examinee i and is commonly referred to as the ability parameter of examinee i; and b_j is the difficulty parameter of item j. Finally, the left hand side of equation 1 is the log odds of $P\left(X_{ij}=1|\,\theta_i,b_j\right)$; that is, the log of the ratio between $P\left(X_{ij}=1|\,\theta_i,b_j\right)$ and $1-P\left(X_{ij}=1|\,\theta_i,b_j\right)$. In IRT, item parameters are commonly treated as fixed effects, which imply that their values are constant across the examinees; whereas the ability is commonly treated as a random effect, which implies that θ_i varies across the examinees. It is common to assume that

$\theta_i \sim N(0, \sigma_\theta^2)$ to resolve the identification issue in IRT models, although other distributions can be assumed as well.

Similarly, the LLTM can be formulated within the GLMM framework as follows:

$$logit\left[P\left(X_{ij}=1 \mid \theta_i, \mathbf{q}_j \boldsymbol{\eta}\right)\right] = \theta_i - \sum_{k=0}^{K} \eta_k q_{jk}\,, \qquad (2)$$

where $\mathbf{q}_j$ $\left(\mathbf{q}_j = \left(q_{j1}, q_{j2}, ..., q_{jK}\right)\right)$ is the vector that contains the values of a set of predictors on item *j* . In the current case, the set of predictors will be the set of identified item features and their values on each item were specified by the researcher as *a priori*; $\boldsymbol{\eta}$ $\left(\boldsymbol{\eta} = \left(\eta_1, \eta_2, ..., \eta_K\right)\right)$ is the vector that contains the estimated effects of the set of predictors on the difficulty of item *j* and they are commonly referred to as the so called *basic parameter*. It is common to set q_{j0} to be 1 so that η_0 becomes the intercept (or a scaling constant).

The Rasch model has an item-specific parameter and does not allow the incorporation of item features into the model. In contrast, the LLTM is capable of modeling the contributions of such variables to item difficulty by decomposing the item parameter b_j into a linear combination of a set of predictors. However, the effects of those predictors are modeled as fixed; that is, values of the basic parameters are assumed to be constant across the examinees. As mentioned earlier, this might be too stringent an assumption for modeling the impacts of item stimulus features on item difficulties. Rijmen and Boeck (2002) proposed a variation of LLTM that relaxed the assumption by allowing the values of some or all of the basic parameters to be differ across examinees. The resulting model is called the random-weights LLTM (hereafter referred to as RW-LLTM) and can be given as follows:

$$logit\left[P\left(X_{ij}=1\,|\,\theta_i,\mathbf{q}_j\boldsymbol{\eta},\boldsymbol{\lambda}_i\right)\right]=\theta_i-\left(\sum_{k=0}^{K_1}\eta_k q_{jk}+\sum_{k=K_1+1}^{K}\lambda_{ik}q_{jk}\right), \quad (3)$$

where K_1 is the number of predictors whose effects are fixed and $K_2\ (=K-K_1)$ is the number of predictors whose effects are modeled as random; $\boldsymbol{\lambda}_i\ \left(\boldsymbol{\lambda}_i=\left(\lambda_{i(K_1+1)},\lambda_{i(K_1+2)},...,\lambda_{iK}\right)\right)$ is the vector of basic parameters for those predictors with random effect. The subscript i in $\boldsymbol{\lambda}_i$ indicates that values of $\boldsymbol{\lambda}_i$ are individual-specific. Because predictors with random weights are in fact predictors whose effects are random, the RW-LLTM is in fact an IRT model with multiple random effects (recall that the examinee ability θ_i is also considered as random). A multivariate normal distribution is commonly assumed for the set of random effects in the model.

The RW-LLTM was one of the two primary models that were used in the current study to model random effects of language components in mathematical items. The other model generalized the RW-LLTM by adding an item-specific error term to the latter. The resulting model can be given as follows:

$$logit\left[P\left(X_{ij}=1\,|\,\theta_i,\mathbf{q}_j\boldsymbol{\eta},\boldsymbol{\lambda}_i\right)\right]=\theta_i-\left(\sum_{k=0}^{K_1}\eta_k q_{jk}+\sum_{k=K_1+1}^{K}\lambda_{ik}q_{jk}+\varepsilon_j\right), \quad (4)$$

where ε_j is the deviate of item j from the predicted value based on the set of predictors and it is assumed that $\varepsilon_j \sim N\left(0,\sigma_\varepsilon^2\right)$. Because ε_j is considered as random across the items, and any other random effects (including θ_i) are considered as random across the examinees, this model is called a crossed random-effects LLTM (hereafter referred to as CRE-LLTM). It is assumed that ε_j is independent of both the predictors (both fixed and random) and examinee ability θ_i.

There is an important reason for adding the error term ε_j in CRE-LLTM. By comparing CRE-LLTM to either LLTM or RW-LLTM, we can investigate the consequence of model misspecification on both estimates and statistical inferences of model parameters because the latter two models make unrealistic model assumptions. Without ε_j and any predictor with random effect in CRE-LLTM, the predicted value of item difficulty for item j from equation 4 is assumed to be equal to the item-specific difficulty value from the Rasch model. It implies that item difficulty can be predicted perfectly by the set of predictors in the model. Obviously, this is a very stringent assumption. It is almost always true that the predicted values of item difficulties are not the same as those from Rasch model given that the number of predictors is fewer than the number of items in the test. Therefore, by adding the item-specific error term ε_j, CRE-LLTM given by equation 4 acknowledges the imperfect predictions of item difficulty parameters from the set of predictors. The estimated value of σ_ε^2 provides an indication of the prediction power of the set of predictors in the model (see Janssen, Schepers & Peres, 2004 for alternative interpretations of ε_j).

Method

Measure and Data

One form of the state mandated test for 3rd grade mathematics test in a Midwest state was chosen as the measure in the current study. The test form consists of 70 multiple-choice items that cover four major domains of the corresponding curriculum: number and computation, algebra, geometry and data analysis. Item response data were collected for 3407 students on the selected form. At the raw score scale, coefficient of alpha is 0.914 for the sample.

Identification of Language Components

The cognitive processing model of mathematical problem solving by Mayer (1985) was used as the theoretical basis for the model-based item analysis. Based on previous research (Embretson, 2004; Embretson and Wetzel, 1987; Shaftel et al, 2006), a set of item stimulus features were identified and their values on each item in the tests were coded. Table 2 presents the set of item stimulus features that were identified as linking to each of the four cognitive processes of mathematical problem solving. According to the theory, impacts of language components on mathematical problem solving primarily occur during the problem translation phase. In particular, two variables were identified as relevant to the complexity of language statements involved in the item: number of words in item stem and proposition density (hereafter refer to as WordStem and PropDnst, respectively). As will be elaborated in later sections, the current study would primarily focus on modeling the effects of these two variables on item difficulty while controlling for other item stimulus features in Table 2.

Modeling Procedure

Following Rijmen and Boeck, a preliminary regression analysis was performed on the logit of the proportion correct values of the set of items in the test using the set of item stimulus features as predictors. The resulting coefficients from the regression model can be used as the basis to compare the estimates from both RW-LLTM and CRE-LLTM.

For all models with random effects, a multivariate normal distribution was assumed for the random effects in the model. The same reparameterization as Rijmen and Boeck did, $\lambda_{ik}^{*} = \lambda_{ik} - \overline{\lambda}_{k}$, was used for parameter λ_{ik} in equation 3 and 4. By doing so, it is assumed that

$\lambda_{ik}^{*} \sim N\left(0, \sigma_{\lambda_{ik}^{*}}^{2}\right)$. Both a random coefficient and a corresponding fixed mean coefficient were estimated for each random effect that is associated with the predictors.

Two sets of IRT models were fit to the data with one set of models based on RW-LLTM and the other set based on CRE-LLTM. For those models within RW-LLTM, the sequence of model fitting for the data is: (a) a LLTM with the ability θ_i as the only random effect; (b) a RW-LLTM with the ability and the variable WordStem as random effects; (c) a RW-LLTM with the ability and the variable PropDnst as random effects ; (d) a RW-LLTM with the ability, WordStem, and PropDnst as random effects, controlling for all other item stimulus features in Table 2. The same sequence of models was also fit to the data using CRE-LLTM, with the exception that an item-specific random error term was added to each of the models. All RW-LLTMs were estimated using SAS NLMIXED procedure, whereas all CRE-LLTM were estimated using the SAS GLIMMIX macro (SAS institute Inc, 2004). The marginal maximum likelihood estimation (MMLE) method was used in SAS NLMIXED procedure. For models with one or two random effects, the number of quadrature points was set to 10. For models with three random effects, five quadrature points was used. To reduce computing time, the estimation was done without the option of adaptive quadrature points. For models of CRE-LLTM, the restricted maximum likelihood estimation (REML) based on the penalized quasi-likelihood (PQL) was used.

Since the primary goal in this study is to model the individual-specific impacts of language components, selection among models mentioned earlier would focus on selection among models with random effects. However, as pointed by Rijmen and Boeck, likelihood ratio test is not appropriate to compare models with different numbers of random effects, even when

such models are nested. Instead, the Bayesian information criterion (BIC; Schwartz, 1988) was used to select among models. A model with the smallest value of BIC was chosen as the favored model. Because the CRE-LLTMs were estimated based on a quasi-likelihood function instead of the actual likelihood function, the corresponding BIC value is not appropriate for model selection, model section was only done among RW-LLTMs using the BIC values instead.

Results

Results from the regression analysis on item logits were presented in Table 3. Collectively the set of item stimulus features explained only 14% of the variance of the logit transformation of the proportion correct difficulties for the 70 items in the test. Except for the intercept, none of the predictors is statistically significant.[1] This might be due to the limited number of items in the test and/or the fact that the set of items is very easy relative to the examinee sample. Proportion correct values for the items ranged from .49 to .96 with a mean of .83. In this case, a negative coefficient indicates that items with high values on the variable are more difficult. Although not statistically significant, all but one item features have the signs of their coefficients consistent with what were expected from the theory. Difficult items tend to have more words, higher proposition density, more number of subgoals, and may require examinee to translate information from a graph/table or generate mathematical equations from the problem statements.

Table 4 presents fit information for the set of models that were fit to the data. Based on the BIC value of each model, the model with both ability and PropDnst as random effects (Model 3) is the model of choice. This model has a BIC value even smaller than that of the model with ability, WordStem, and PropDnst as random effects (the corresponding values of BIC are 224860 and 225070, respectively). The estimated variance-covariance matrix for each

of the four RW-LLTM models was also presented in Table 4. Using the estimates of model 4, the estimated variances were 1.029, .0123 and 2.907, respectively, for the ability, number of words in stem and proposition density. The effects of both number of words in stem and proposition density were somewhat negatively correlated with the examinee abilities. The correlation coefficient between number of words in stem and ability is -.231. The correlation coefficient between proposition density and ability is -.352. The effects between the two language components were positively correlated with a coefficient of .213. This suggests that there is a substantial variation across examinees in terms of the difficulty of items with high proposition density, which represents a distinctive aspect of variation other than the variation among examinee abilities. In general, items with high proposition density tend to be more difficult (see Table 3), but this aspect of item has more impact on examinees with low abilities than those with high abilities. The large variation in the impact of the proposition density was also found in the other model that contains this parameter. Similarly, items with large number of words tend to be more difficult, but its impact varies across examinees. However, compare to proposition density, the variation among examinees in terms of the difficulty of items with a lot of words was relatively small. This small variation may explain why the model with ability and proposition density as random effects (model 3) has a smaller BIC value than the model that adds an extra random effect of the WordStem variable (model 4) as the former is more parsimonious.

Parameter estimates for the fixed effects and the means of the random effects were presented in Table 3. Because of the model formulation given in equation 1 to 4, signs of parameters from both RW-LLTM and CRE-LLTM are different from those from the multiple regression analyses. A positive coefficient from RW-LLTM and CRE-LLTM indicates that the items with higher values on the variable are more difficult. To facilitate comparison of parameter

estimates across models, estimates from RW-LLTM and CRE-LLTM were multiplied by -1. In term of absolute magnitude, both multiple regression analysis on item logits and the RW-LLTM on item response yielded similar values on the set of item stimulus features. However, all except one fixed effect estimated from the RW-LLTM are statistically significant now. There are two possible reasons for this. In RW-LLTM, test of significance for a parameter is based on a much larger degree of freedom than that in multiple regression analysis on item logits. Second, RW-LLTM makes a strong assumption that the variables perfectly predict item difficulty, which is clearly not true in this case. Comparison between RW-LLTM and CRE-LLTM with the same model structure might provide us a clear answer to this question.

In general, the estimated variances and covariances are similar in value between RW-LLTM and CRE-LLTM, although estimates from CRE-LLTM tend to be slightly larger (see Table 4). Using the set of item stimulus features as predictors, the estimated item error variance of the chose model 3 was .726, which is somewhat consistent with what we estimated from the multiple regression analysis on item logits. For fixed effects, estimates of CRE-LLTM are similar in values to those of multiple regression model and RW-LLTM (see Table 3). Different from results of RW-LLTM, however, inferential tests for the fixed effects yielded by CRE-LLTM are consistent with those tests by the multiple regression model. Since the corresponding model in RW-LLTM and CRE-LLTM differ only by the item error term, test of significance for a parameter in those two models has similar degree of freedom. This finding, along with the consistency of the item error variance, suggests that inferential tests for estimates of fixed effects from RW-LLTM may be suspected, even though the magnitudes of such estimates are similar to those from CRE-LLTM and the multiple regression model. As mentioned earlier, this might be

due to the strong assumption that RW-LLTM made, which consequently affect the standard error estimate associated with each of the parameters in the model.

Summary and Discussion

Impacts of language components in mathematical word problems on examinee performances were examined in this study using IRT models with multiple random effects. Based on a cognitive theory of mathematical problem solving and previous studies on text comprehension, a set of item stimulus features were identified as linking to different cognitive processes of mathematical problem solving. In particular, two aspects of language statements were identified as linking to the phase of problem translation in mathematical items: number of word in item stem and proposition density. Results from the study suggested that, although the average impact of language components on mathematical problem solving is not significant for the examinee sample, substantial variations among examinees of the impacts from such language components were found for both variables. Specifically, there is a relatively large variation across examinees in terms of the difficulty of items with high proposition density, which represents a distinctive aspect of variation other than the variation among examinee abilities. In general, items with high proposition density tend to be more difficult, but this aspect of item has more impact on examinees with low abilities than those with high abilities. Similarly, items with large number of words tend to be more difficult, but its impact varies across examinees. However, compare to proposition density, the variation among examinees in terms of the difficulty of items with a lot of words was relatively small.

As mentioned earlier, previous studies primarily fall into one of three different tracks: (1) examining the relationship between reading abilities (general or specific) and mathematical abilities of examinees; (2) whether a general language factor can be identified through factor

analysis based on the test data; and (3) examining the average impacts of the specific item features on item properties through either multiple regression models or LLTM. In contrast, the approach using IRT models with multiple random effect to model impact of language components on mathematical items is unique in that it provides a method to examine not only the average impacts of the specific item features on item properties over the sample, but also variations of such impacts of those item features among the examinees in the sample. Consequently, it allows researchers to identify various sources of individual differences other than the general abilities of examinees. Results from this study showed that individual-specific effects of item features on item properties represented indeed distinctive aspects of individual differences in the data. This has important implications for the appropriate and adequate inferences and interpretations of scores from the tests. While knowledge about the average effects of item stimulus features facilitate item and test development, knowledge about the individual-specific effects of such features help appropriate interpretations of examinees scores.

Two sets of models were used in the current study to model the impacts of language components in mathematical items: the RW-LLTM and the CRE-LLTM. The major difference between the two models is that the latter adds an item-specific error term to the former. By doing so, the CRE-LLTM acknowledges the fact that it is rare in practice that a set of identified predictors can perfectly predict item difficulties. Therefore, the CRE-LLTM allows the researchers to model not only the individual-specific effects of some predictors, but also the adequacy of such a set of predictors on predicting item properties. It turns out that the allowance of an item-specific error term in the model has a significant consequence on inferential test of the resulting parameter estimates. Results from the study suggested that inferential tests of parameter estimates from RW-LLTM were inflated (estimates of parameters were more likely to be

significant), although the values of such estimates are comparable to those from CRE-LLTM. This might be due to the fact that the RW-LLTM makes a stringent assumption about the adequacy of the set of item predictors in the model, which leads to distorted estimates of the standard errors that are associated with different parameter estimates. In contrast, inferential tests of parameter estimates from CRE-LLTM are more sensible and consistent with those from the multiple regression analysis. It should be pointed out that the statements here are only speculative in that the true values of such parameters were unknown. Simulation study needs to be done in order to provide a solid answer to this question.

In conclusion, assessing quantitative reasoning is one of the major topics in educational setting. Important decisions are routinely made for individual competence, instructional intervention as well as policy based on results of such assessments. Therefore, a clear understanding of the impacts of language factors in mathematical items on the process of problem solving of examinees will no doubt contribute to the appropriate and adequate inference and interpretation of scores from such tests (Messick, 1995). It is the hope of the current study to enhance our capability of capturing such impacts through appropriate approach to data analysis.

Author's Note

[1]For technical reason, the variable WordStem was rescaled. This was done through dividing each value on the variable by 10. So the estimated coefficient on this variable corresponds to the effect on item difficulty of every 10 words increase in an item. The rescaled variable was used throughout all of the analyses in this study.

References

Abedi J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219 – 234.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*, 359 – 385.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612 -637.

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

DeGuire, L. J. (1985). *The structure of mathematical abilities: the view from factor analysis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: application to mathematical reasoning. *Journal of Educational Measurement, 32*, 277 – 294.

Embretson, S. E. (2004). *A cognitive model of mathematical reasoning*. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology. Naples, FL: October.

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

Enright, M. K., & Sheehan, K. M. (2001). Modeling the difficulty of quantitative reasoning items: implications for item generation. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development, (pp. 129 -158).* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6,* 223 – 283.

Hanson, M. R., Hayes, J. R., Schriver, K., LeMahieu, P. G., & Brown, P. J. (1998). *A plain language approach to the revision of test items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Jassen, 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, NY: Springer.

Johnson, J. T. (1949). On the nature of problem solving in arithmetic. *Journal of Educational research, 43,* 110 -115.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29,* 35 -45.

Jonassen, D. H. (2003). Designing research-based instruction for story problems. *Educational Psychology Review, 15,* 267 -296.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109 – 129.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363 -394.

Kopriva, R.J (1999). *Ensuring accuracy in testing for English language learners: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Lucangeli, D., Tressoldi, P., & Cendron, M. (2002). Cognitive and metacognitive abilities involved in the solution of mathematical word problems: validation of a comprehensive model. Contemporary Educational Psychology, 23, 3, 257 – 275.

Mayer, R.E. (1985). Mathematical ability. In R.J. Sternberg (Ed.), *Human abilities: Information processing approach* (pp. 127-150). San Francisco, CA: Freeman.

Mayer, R. E. (1987). *Educational psychology: A cognitive approach,* (p.345-373). Boston: Little, Brown.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist,* 50, 9, 741-749.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rijmen, F., & Boeck, P. D. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26,* 271 – 285.

Rijmen, F., Tuerlinckx, F., Boeck, P. D., & Kuppens, P. (2003). *A nonlinear mixed model framework for item response theory. Psychological Methods, 8,* 185 -205.

SAS Institute Inc. (2004). *SAS online doc 9.13*. Cary,NC: SAS Institute Inc.

Schoenfeld, A. H. (1983). Episodes and executive decisions in mathematical problem solving. In R. Lesh & M. Landau, M. (Eds.). *Acquisition of mathematics concepts and processes (pp. 345 -395)*. New York, NY: Academic.

Schwarz, G. (1988). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461 -464.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using Algebra word problem to assess quantitative ability: Attributes, Strategies, and Errors. *Cognition and Instruction, 14,* 285 -343.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and student with disabilities. *Educational Assessment, 11,* 105 – 126.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp 263 – 331). New York, NY: American Council on Education /Macmillan.

Wu, M., & Adams, R. (2006). Modeling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18,* 93 -113.

Table 1

Proporsitions in the Sentence Example

Proposition number	Proposition
1	(Math, Teacher)
2	(and, John, Mary)
3	(liked, P2,Teacher)

Table 2

Definitions of Cognitive Processes through Item Characteristics

Stage/Variable	Detailed Description
Problem Translation	
Wordstem	Number of words in item stem
PropDnst	Proposition density in an item
Problem Integration	
Trangrph	Translate graph or diagram (1=yes, 0 =no)
GenEqu	Generate or translate equation (1=yes,0=no)
Solution Planning	
NumGoals	Number of goals or subgoals in an item
Solution Execution	
Computation	Number of computation required

Table 3

Parameter Estimates for Fixed Effects from Different Models

Parameter	Multiple Regression			RW-LLTM			CRE-LLTM		
	Estimate	SE	t	Estimate*	SE	t	Estimate*	SE	t
Intercept	2.432	0.327	7.436	2.576	0.029	89.778	2.961	0.431	6.868
WordStem	-0.103	0.085	-1.212	-0.098	0.006	-15.856	-0.093	0.111	-0.842
PropDnst	-0.990	0.733	-1.351	-0.852	0.061	-14.084	-1.135	0.962	-1.181
Trangrph	-0.222	0.183	-1.217	-0.313	0.012	-25.785	-0.291	0.238	-1.225
GenEqu	-0.310	0.207	-1.493	-0.368	0.014	-27.032	-0.403	0.270	-1.494
NumGoals	-0.019	0.133	-0.143	-0.006	0.008	-0.738	-0.053	0.172	-0.310
Comput	0.002	0.034	0.048	0.017	0.002	7.889	-0.004	0.044	-0.102

Note: *Estimates were multiplied by -1 to facilitate comparison among models.

Table 4

Fit Indices and Variance-Covariance Estimates from RW-LLTMs and CRE-LLTMs

Model	BIC	RW-LLTM			CRE-LLTM			
		Ability	WordStem	PropDnst	Ability	WordStem	PropDnst	Item
Model 1	225155							
Ability		0.860	--	--	1.144	--	--	
WordStem			--	--		--	--	
PropDnst				--			--	
Intercept (Item)								0.715
Model 2	225092							
Ability		0.982	-0.023	--	1.237	-0.038	--	
WordStem		***-0.247***	0.009	--	***-0.272***	0.016	--	
PropDnst				--			--	
Intercept (Item)								0.717
Model 3	224860							
Ability		1.071	--	-0.594	1.447	--	-1.259	
WordStem			--	--		--	--	
PropDnst		***-0.357***		2.581	***-0.463***		5.105	
Intercept (Item)								0.726
Model 4	225070							
Ability		1.029	-0.026	-0.609	1.527	-0.039	-1.074	
WordStem		***-0.231***	0.012	0.040	***-0.247***	0.017	0.011	
PropDnst		***-0.352***	***0.213***	2.907	***-0.397***	***0.037***	4.792	
Intercept (Item)								0.808

Note: Number in bold indicates correlation coefficient.