Asia Pacific Education Review

An IRT Modeling of Change over Time for Repeated Measures Item Response Data Using a Random Weights Linear Logistic Test Model Approach

--Manuscript Draft--

Manuscript Number:	
Full Title:	An IRT Modeling of Change over Time for Repeated Measures Item Response Data Using a Random Weights Linear Logistic Test Model Approach
Article Type:	Original Article
Corresponding Author:	Insu Paek Florida State Univeristy Tallahassee, Florida UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Florida State Univeristy
Corresponding Author's Secondary Institution:	
First Author:	Insu Paek
First Author Secondary Information:	
All Authors:	Insu Paek
	Sun-Geun Baek, Ph.D.
	Mark Wilson, Ph.D.
All Authors Secondary Information:	
Abstract:	In this study, repeated measures item response data from an English language proficiency (ELP) test administered over two years were analyzed by item response theory (IRT)-based approaches to model change over time. First, change over time was modeled by an item side change approach where change was parameterized at the item level. Second, more importantly, this study introduced and demonstrated an application of an interaction model for change which overcomes a limitation of the item side change approach by allowing interaction effects between persons and item side change parameters. The results showed that student changes in the ELP data used in this study should be modeled for the three sub-content areas and that modeling interaction for change was necessary.
Suggested Reviewers:	

Title Page

An IRT Modeling of Change over Time for Repeated Measures Item Response Data Using a Random Weights Linear Logistic Test Model Approach

Insu Paek

Florida State University

Sun-Geun Baek

Seoul National University

Mark Wilson

University of California at Berkeley

Insu Paek, Assistant Professor
Educational Psychology & Learning System
Florida State University
3204D Stone Building
1114 W. Call St.
Tallahassee, FL 32306-4453
USA
(Tel) 850-644-3064
Email: ipaek@fsu.edu

Sun-Geun Baek, Professor
Department of Education
Seoul National University
San 56-1 Byunji
Sillim 9-dong, Gwank-gu
Seoul, 151-748
South Korea
(Tel) 82-2-880-7645
Email: dr100@snu.ac.kr

Mark Wilson, Professor
Graduate School of Education
4415 Tolman Hall
University of California, Berkeley
Berkeley, CA 94720
Email: markw@berkeley.edu

An IRT Modeling of Change over Time for Repeated Measures Item Response Data Using a Random Weights Linear Logistic Test Model Approach

Abstract

In this study, repeated measures item response data from an English language proficiency (ELP) test administered over two years were analyzed by item response theory (IRT)-based approaches to model change over time. First, change over time was modeled by an item side change approach where change was parameterized at the item level. Second, more importantly, this study introduced and demonstrated an application of an interaction model for change which overcomes a limitation of the item side change approach by allowing interaction effects between persons and item side change parameters. The results showed that student changes in the ELP data used in this study should be modeled for the three sub-content areas and that modeling interaction for change was necessary.

Keywords: IRT for Repeated measures data, Longitudinal data analysis

Introduction

Tracking subjects' change over time is an important matter in educational achievement testing. The No Child Left Behind Act (NCLB) in 2001 has spurred an interest in measuring change. Currently the kind of change that is getting measured, most often, is based on cross-sectional student data. For instance, the average score on a test for the fifth graders in 2005 is compared to the average for a new cohort of the fifth graders in 2006. But at the same time, many states, with NCLB, are getting interested in collecting longitudinal data for student cohorts over time, and using this as the basis for measuring change. It seems that what is happening in state achievement test programs with NCLB provides an entry point for a discussion or use of item response theory (IRT) based approaches to modeling change (e.g., Roberts and Ma, 2005).

There are many different methodologies that can be used for modeling change with repeated measures data. (Comparison of the different methodologies with the IRT modeling for change was not made in this paper, which is beyond the current capacity of this paper.) In the IRT literature, several IRT-based approaches to modeling change have been also shown (Embretson, 1991; Fischer, 1989, 1995b; Hoijtink, 1995; Robert & Ma, 2005**;** te Marvelde, Glas, Van Landeghem, & Van Damme, 2006; Wang & Chyi-In, 2004; Wang, Wilson, & Adams, 1998). Perhaps the most distinctive feature in the IRT modeling for change is that IRT modeling considers discrete nature of data (dichotomous item responses here) and allows for optimal approximations to interval level measurement with the incorporation of measurement error (Fischer, 1995a; Wright & Masters, 1982), when the model describes the data reasonably well.

Relationship between the observed raw score scale and the latent scale created from the IRT modeling is nonlinear. The same amount of differences in two different pairs of the observed raw scores does not have the same meaning in the latent scale. This can lead to different interpretations of change patterns in the data, contingent upon which scores (the IRT score or the observed raw score) are used for investigation. Another distinctive feature in the IRT modeling for change is, because the basic unit of modeling in IRT is item level, that it allows the modeling of change as a change in item characteristic (e.g., item difficulty), so that the overall change or trend can be decomposed into change in individual item level.

In this study, we introduce and fit several IRT change models into repeated measures data. We start with an IRT change modeling where the change is modeled with regard to item side characteristics, building a best change model for the repeated measures data from English language proficiency (ELP) test administrations over two years. The item side change modeling approach is basically a main effect model because change is parameterized as item parameters which are constant across subjects. Therefore, it may ignore potential interactions between item and person side characteristics in modeling change. For this reason, in this study, we extend the item side change modeling to an approach where interactions between person and item sides are taken into account in the IRT change modeling. This IRT interaction modeling approach for change was never used for assessment of change and does not yet appear in the IRT change modeling literature either. The explanations of the IRT change models which were used in this study are given in detail in a later section, followed by the estimation results from the item side change approach and the interaction modeling approach.

Data

The data were dichotomous responses to a large scale assessment of English language proficiency (ELP) in the United States. A group of students in grade 4 took the ELP test in 2004, and then took the same test again the next year in grade 5. The data used in this study had complete item responses from 248 students for both time points. The ELP test had five sub-content areas: Listening, Writing Conventions, Reading, Writing, and Speaking. In what follows we restricted our analysis to the areas of Listening, Writing Conventions, and Reading. We examined responses to 20 items for each sub-content area, for a total of 60 items.

IRT Change Models

The starting point for any IRT change model is a set of assumptions about the distribution of discrete item responses. In common with other IRT models, we assume the conditional distribution of an item response given person ability is a Bernoulli distribution with probability of a correct answer equal to P_i, where the subscript i represents items, ranging from 1 to I. The general form of the IRT change models employed in this study has P_i as follows

$$P_i = \frac{\exp(\eta)}{1+\exp(\eta)}, \tag{1}$$

where η has an additive linear form regarding parameters for person, item, and change parameters. For example, η may consists of person ability parameter, item parameter

and item side change parameter, or it may have person, item, change parameters, and an interaction parameter between person and item sides. Because a linear additive parameterization is used in η, the IRT change models here have a close form to the linear logistic test model (LLTM: Fischer, 1973). All item side change models in this study can be considered as special cases of LLTM, but the extension we make here, i.e, the interaction modeling approach for change cannot be subsumed under LLTM. It turns out that the interaction model for change can be formulated as a special case of the random weights linear logistic test model (RWLLTM: Rijmen & De Boeck, 2002) which is an extension of LLTM. In this sense, all IRT change models in this study are within the framework of RWLLTM.

Item Side Change Modeling Approach

Typically, change is thought of as person ability change (, which is true given the same stimuli (items) were administered), but without loss of generality, the change can be expressed with respect to item difficulty side. When the change is modeled in the item side and show how much change occurs in each item, it would be of use for instructional purpose as well. The item side change models that were tried in this study are as follows.

- M1: Model with no change between time points 1 (Year 2004) and 2 (Year 2005).

$$\eta = \theta - \delta_{i(\mathrm{T1=T2})}\,, \tag{2}$$

where θ is person ability and $\delta_{i(\mathrm{T1=T2})}$ (i = 1,2, …60) is the item difficulty which is the same between time points 1 and 2.

- M2: Overall change model. This is the simplest summary of the change and all changes are summarized as a constant.

$$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{O}, \quad (3)$$

where $\delta_{i\mathrm{T1}}$ is item difficulty at time point 1, and δ_{O} is an overall change parameter between time points 1 and 2. .

- M3: Sub-content area change model. Given that the ELP test has three sub-content areas, the model that captures change in each sub-content area is:

$$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{S_j}, \quad (4)$$

where δ_{Sj} (j = 1, 2, 3) is the change in each sub-content area (Listening, Writing Conventions, and Reading, respectively) between time points 1 and 2.

- M4: Item-by-item change model. This is the most general model in the item side change approach.

$$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{i\mathrm{C}}, \quad (5)$$

where, again, $\delta_{i\mathrm{T1}}$ is item difficulty at time point 1 and $\delta_{i\mathrm{C}}$ is the item difficulty change parameter, i.e., $\delta_{i\mathrm{C}} = \delta_{i\mathrm{T1}} - \delta_{i\mathrm{T2}}$ ($\delta_{i\mathrm{T2}}$ is item difficulty at time point 2). Note that the item-by-item difficulty change and its standard error are directly estimated from the data, thereby a test for $\mathrm{H_o}$: $\delta_{i\mathrm{C}} = 0$ can be easily conducted by the z-test (or Wald test), $z = \frac{\hat{\delta}_{iC}}{SE(\hat{\delta}_{iC})}$ where $\hat{\delta}_{iC}$ is the maximum likelihood estimate of δ_{iC} and $SE(\hat{\delta}_{iC})$ is the standard error.

- M5: Partial item-by-item change model. After fitting the item-by-item change model and performing a test for $\mathrm{H_o}$: $\delta_{i\mathrm{C}} = 0$, a sub-model where the non-significant $\delta_{i\mathrm{C}}$ is fixed as zero can be fitted as a more parsimonious way to describe the data than the item-by-item change model (M4).. The Partial item-by-item change model is

$$\eta = \theta - \delta_{i\mathrm{T1}} + \delta^{*}_{i\mathrm{C}}, \qquad (6)$$

where $\delta^{*}_{i\mathrm{C}}$ is the difficulty change parameter for which its $\delta_{i\mathrm{C}}$ in M4 is statistically significant by the z-test under $\alpha = 0.05$.

M1 through M4 are hierarchically structured. M1 is nested within M2. M2 is nested within M3, and M3 is nested within M4. M5 is nested within M4 while M1 is nested within M5, but M2 and M3 are not necessarily nested within M5. (It turns out from the z test results for H_o: $\delta_{i\mathrm{C}} = 0$ that M2 and M3 are not nested within M5.) The parameter θ is a random parameter and all the other parameters are fixed unknown parameters in all models in the item side change modeling approach. The random parameter θ was assumed to follow a normal distribution.

For hierarchically nested models, the log-likelihood ratio (LR) test was conducted to compare a reduced model with a general model. Statistical significance for the LR test means that the reduced model is rejected, favoring the general model. In addition, two descriptive model fit indices were computed. They are the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978), which are applicable when the maximum likelihood estimation of the model parameters is conducted. Both AIC and BIC penalize a model by taking the number of parameters in a model into account for computing their model-data fit indices. AIC and BIC can be used for nested and non-nested model fit comparisons. A model with a small AIC (or BIC) is preferred. The two indices do not necessarily coincide with each other in their best model selection and BIC tends to select simpler models than AIC (Lin & Dayton, 1997).

Interaction Modeling Approach for Change

So far, changes were modeled through the item (difficulty) side. Another approach is to consider the interaction between persons and the changes in the item side. In this approach, the change in the item side depends on each person, i.e., the interaction model for change permits person-dependent effect of item side change parameters. The interaction model for change was applied to extend the best fit model of the item side change approach, which was M3, the sub-content area change model.

- M6: Interaction model for change

$$\eta = \theta_n - \delta_{i\mathrm{T1}} + \delta_{S_J} + \tau_{nS_J}, \quad (7)$$

where τ_{nS_J} is the interaction term between θ_{n} and δ_{S_J}, i.e., the sub-content area change effect varies, depending on each person. The person parameter θ_{n} and the interaction parameter τ_{nS_J} are random parameters with other parameters as unknown fixed parameters as before. τ_{nS_J} is estimated as a deviation from δ_{S_J} that is the average change for a given j. θ_{n} and τ_{nS_J} were modeled to take a multivariate normal distribution in this study. This model provides average changes for each sub-content area between the two time points by δ_{S_J} and the extent of the interaction between persons and the sub-content area change by the variance estimate of τ_{nS_J}. The sub-content area change model (M3) is nested within the interaction model for change (M6). When the variance of τ_{nS_J} is zero, M6 is reduced to M3. AIC and BIC were used for the model comparisons, but the LR test was not conducted because the variance equal to zero in the null hypothesis puts the reduced model on the boundary of the general (alternative) model parameter space. In that case, the regular LR test is not applicable.

Estimation of the Models

All IRT change models in this study were estimated by the program, ConQuest (Wu, Adams, & Wilson, 1998). A generalized linear and nonlinear approach (e.g., using SAS NLMIXED [SAS Institute, 1999] procedure) may be taken to estimate all the models employed in this paper as well (See, e.g., De Boeck & Wilson, 2004). Suppose that the data matrix is **X** and let $H(\cdot) = H(\theta \mid \mu, \Sigma)$ be the distribution function of θ, where θ is a vector of random parameters which follow a multivariate normal distribution, μ, is a mean vector, and Σ is a covariance matrix . Then the marginal likelihood function with conditional independence assumption is:

$$L(\delta, \mu, \Sigma \mid \mathbf{X}) = \prod_n \int \prod_i P_i^{x_i} (1 - P_i)^{1-x_i} dH ,$$

where n is the person index ($n = 1,2,\ldots N$) and P_i is as defined as before in Equation 1. ConQuest estimates the parameter values maximizing the above likelihood by the Expectation-Maximization algorithm (Bock & Aitken, 1981).

Result

Before proceeding with results, item fit was investigated using weighted mean square and weighted t statistics (Wright & Masters, 1982; Wu, 1997) to examine the extent of item fit to a linear additive IRT model which has a constant slope across items, i.e., Rasch type modeling. The weighted t and the weighted mean squared statistics were provided by the ConQuest program. The mean of the weighted mean square becomes close to one when item fits the model well and the weighted t is considered to be an

approximate normal deviate. These item fit statistics were from fitting the Rasch model to the time point 1 data and the time point 2 data sets, respectively. Wilson (2004) suggested the use of both item statistics. Item was considered as misfit if |weighted t | > 2 and if weighted mean square < 0.75 or weighted mean square > 1.33. These weighted mean square criteria were suggested by Adams and Khoo (1996) and Wilson (2004). Under both the weighted t and the weighted mean square criteria above, no items were declared as misfit items.

Item Side Change Modeling Approach

Since the item-by-item change model (M4) was the most general model in the item side change approach, first, we show its results in Figure 1. The black dots in Figure 1 represent the item difficulty change estimates [$\delta_{iC} = \delta_{iT1} - \delta_{iT2}$: time point 1 (Year 2004) – item point 2 (Year 2005)].

Figure 1

All item changes but three (items 1, 2, and 23) were positive, indicating that students showed positive change. To investigate whether the changes were significant, the z-test was conducted with $\alpha = .05$. A total of 13 item changes were non-significant. The Listening section had the largest non-significant number of item changes among the three sub-content areas: 9 non-significant items (items 1, 2, 3, 4, 11, 14, 18, 19, and 20). Writing Conventions had 2 non-significant items (items 23 and 24). Reading had also 2 non-significant items (items 42 and 55).

Comparisons of the models by AIC, BIC, and the LR test are summarized in Tables 1 and 2. In Table 1, in addition to AIC and BIC, the number of parameters in a model and the model deviance which is -2×logliklihood of the model are also reported. (Deviance always becomes smaller when more parameters are used in a model.)

Tables 1 and 2

AIC and BIC identified either the overall change (M2) or the sub-content area change model (M3) as the best model (see Table 1). The model with no change (M1) was the least preferred model by both AIC and BIC. When M1 was compared with the other change models by the LR test, the other change models were always preferred (see Table 2). These indicate that significant change took place and a parsimonious model which explains the data as good as the most general item-by-item change model may be found. The partial item-by-item model (M5) was preferred to the item-by-item change model (M4) by the LR test, AIC and BIC. Overall and sub-content area change models (M2 and M3) are not nested within the partial item-by-item change model (M5). Using the AIC and BIC criteria, both M2 and M3 were preferred to M5. The LR test showed that the sub-content area change model (M3) was preferred to the overall change model (M2), but AIC and BIC disagreed when M2 and M3 were compared. M3 was preferred to M2 by AIC while M2 was preferred to M3 by BIC. As noted before, BIC usually tends to select a simpler model than AIC. When looking at the actual BIC values (M2 BIC = 22647.38 and M3 BIC = 22647.58), their difference was 0.2 which is very small. Given this very small difference in the BIC value, the preference for M3 over M2 by the LR test, and the

AIC criterion result, the sub-content area change model (M3) was considered the best model in the item side change approach, i.e., the change which took place in the ELP data can be summarized parsimoniously through the change in each sub-content area.

The change estimates in the sub-content area change model were .50 (.05), .65 (.04), and .78 (.04) for Listening, Writing Conventions, and Reading, respectively. (The values in the parentheses are standard errors.). All these sub-content area changes were statistically significant when the z-test was conducted. All changes were positive, indicating that students made significant positive changes in each of the three sub-content areas. The Reading section showed the largest change, Writing Conventions the second largest, and Listening the least amount of change among the three sub-content areas.

Interaction Modeling Approach for Change

The interaction model for change (M6) inherits all model properties and interpretations from the sub-content area change model (M3) and in addition, the interaction effect between persons and the each sub-content area change was estimated. In the interaction modeling for change (see Equation 7), the θ_n signifies the person latent score at time point 1 (overall latent score θ at time point 1) and τ_{nS_J} is the interaction term that is a deviation from δ_{S_J} which is the average change for each sub-content area. The results for the interaction model for change are shown in Table 3.

Table 3

The estimated average changes for sub-content areas were 0.42, 0.72, and 0.90 for Listening, Writing Conventions, and Reading, respectively. The amount of changes were

not the same as those of the sub-content area change model (M3), but the order of the sub-content areas regarding the relative amount of changes was the same, i.e, Reading, Writing Conventions and Listening in the order of largest changes as in the sub-content area change model (M3). Again each sub-content area change, estimated by the interaction model for change in Table 3, was statistically significant when the z-test was conducted.

The variance estimates of the interactions (var(τ_{nS_j}) = 0.34, 0.35, and 0.39 for Listening, Writing Conventions, and Reading, respectively) in Table 3 indicate some interactions between persons and the change for each content area. This observation of extant interaction was supported by AIC and BIC as well. The interaction model for change was preferred to the sub-content area change model by both AIC and BIC criteria. (The AIC and the BIC values for the interaction model for change were smaller than those for the sub-content area change model. See Tables 1 and 3.) Note again that although these two models were hierarchically nested, the LR test was not conducted because the reduction of the random parameter (i.e., variance equal to zero) was the null hypothesis of the LR test.

The latent correlation structures among the sub-content area changes in Table 3 showed all positive relationships (.40, .29, and .58 for pairwise correlations among Listening, Writing Conventions, and Reading). The change in Reading had the highest association (.58) with that in Writing Conventions, i.e., a large positive change in Reading tends to be related to a large positive change in Writing Conventions. The relationship between Listening and Reading was not quite strong (.29). With respect to the relationship between the overall latent score (θ) at time point 1 and each sub-content

area change, Listening showed -0.60 correlation, meaning that high overall ability students at time point 1 tend to show smaller positive change in Listening than the low overall ability students at time point 1. Writing Conventions had -0.13 correlation and Reading had 0.12. When a statistical testing of zero correlation using a typical large sample based formula, $z = r\sqrt{N-1}$ where r is the correlation and N is the sample size, was conducted for these three correlations, Listening and Writing Conventions were significant under $\alpha = 0.05$.

Summary and Discussion

Modeling change over time is receiving more attention with NCLB these years. This study introduced and fitted a series of IRT-based item side change models for repeated measures data from an ELP test administered over two years. More importantly, this study introduced an IRT change model which allows for interactions between item and person side characteristics. In the item side change modeling approach, changes are modeled by item level characteristic parameters and their estimated effects are constant across persons. A natural extension of this main effect approach for change modeling is to consider interactions between persons and item side change parameters, thereby permitting person-dependent effects for change parameters in the item side change modeling. The application of the IRT-based interaction modeling for change, demonstrated in this study, was an attempt to overcome a limitation of the IRT-based item side change modeling for change over time in repeated measures data.

The result from the interaction model for change indicated that the interaction effect is not ignorable in the ELP data used in this study and that the sub-content areas in the order of the amount of positive change were Reading, Writing Conventions, and Listening from the largest to the smallest. It is not yet clear that this observation would hold in other data from English language proficiency tests. The latent correlations (see Table 3) show that Writing Conventions and Reading has the strongest association of all pairwise correlations among the three sub-content areas while Listening tends to show lower correlations with Reading and Writing Conventions. Based on these, one might conjecture that when learning English, listening takes a slow and different learning pattern compared to Reading and Writing Conventions. More studies with wider range of longitudinal data will be required to reach general conclusions about systematic patterns in change for the sub-content areas.

One limitation of the IRT-based change models presented in this study is that it depends upon a repeated measures design in which the same items are presented multiple times. For tests administered longitudinally with a mixture of unique and common items, the approaches shown here can only be applied to the common items, which limits applicability of these approaches unless the number of common items is relatively large and they have been properly selected to be representative of the construct of measurement. When repeated measures data are available with appropriate auxiliary information variables such as student background variables, an extension of the interaction model for change (M6) is possible. Once the interaction is found to be significant enough and person level (observed) variables of interest are available, an explanatory version of the interaction model for change may be employed. One

straightforward version of the explanatory interaction model (, which was not done in this study,) will have the following form:

$$\eta = \theta_n - \delta_{i\mathrm{T1}} + \delta_{Sj} + \tau_{nS_J},$$

$$\tau_{nS_J} = Z'\xi_{nj} + \varphi_{nj}, \text{ and}$$

$$\varphi_{nj} \sim \mathrm{N}(0, \omega_j^2)$$

where Z' is a row vector of observed predictors for a person n and a given j, ξ_{nj} is a vector of fixed unknown regression coefficient, and φ_{nj} is a random effect for n and j. This explanatory interaction model for change provides explanation of the variation found across person n for a given j. Lastly it is noted that the change models used in this study can be straightforwardly extended to accommodate more than two time points and polytomously scored item responses.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest.* Melbourne, Australia: Australia Council for Educational Research.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443-459.

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

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning

and change. *Psychometrika, 56*, 495-515

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acto Psychologica, 37*, 357-374.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599-624.

Fischer, G. H. (1995a). Derivations of the Rasch models. In G. H. Fischer, & I. W. Molenaar. (Eds.), *Rasch models: Foundations, recent developments, and applications (pp. 15-38).* NY: Springer-Verg.

Fischer, G. H. (1995b). Linear logistic models for change. In G. H. Fischer, & I. W. Molenaar. (Eds.), *Rasch models: Foundations, recent developments, and applications (pp. 157-180).* NY: Springer-Verg.

Hoijtink, H. (1995). Linear and repeated measures models for the person parameters. In G. H.Fischer, & I. W. Molenaar. (Eds.), *Rasch models: Foundations, recent developments, and applications (pp. 204-214).* NY: Springer-Verg.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249-264.

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

Robert, J. S., & Ma, Q. (2005). IRT models to assess change across repeated measurements. Presentation at the university of Maryland conference.

SAS Institute (1999). *SAS online doc. (version 8).* Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,*

461-464.

te Marvelde, J. M., Glas, C. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66*, 5-34.

Wang, W., & Chyi-In. (2004). Gain scores in item response theory as an effect size measure. *Educational and Psychological Measurement, 64*, 758-780.

Wang, W., Wilson, M., & Adams, R.J. (1998). Measuring individual differences in change with Rasch models. *Journal of Outcome Measurement, 2*(3), 240-265.

Wilson, M. (2004*). Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models*. Unpublished masters dissertation. University of Melbourne.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software*. Australia: ACER.

Table 1. Model Comparisons by AIC and BIC for Item Side Change Modeling Approach

Model	Equation	Deviance	No. of par	AIC	BIC
			AIC order		
M3: Sub-content area change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{S_J}$	22294.72	64	22422.72	22647.58
M2: Overall change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{O}$	22305.55	62	22429.55	22647.38
M5: Partial item-by-item change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta^{*}_{i\mathrm{C}}$	22252.51	108	22468.51	22847.96
M4: Item-by-item change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{i\mathrm{C}}$	22239.02	121	22481.02	22906.14
M1: No change	$\eta = \theta - \delta_{i(\mathrm{T1=T2})}$	22673.46	61	22795.46	23009.78
			BIC order		
M2: Overall change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{O}$	22305.55	62	22429.55	22647.38
M3: Sub-content area change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{S_J}$	22294.72	64	22422.72	22647.58
M5: Partial item-by-item change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta^{*}_{i\mathrm{C}}$	22252.51	108	22468.51	22847.96
M4: Item-by-item change	$\eta = \theta - \delta_{i\mathrm{T1}} + \delta_{i\mathrm{C}}$	22239.02	121	22481.02	22906.14
M1: No change	$\eta = \theta - \delta_{i(\mathrm{T1=T2})}$	22673.46	61	22795.46	23009.78

Table 2. Model Comparisons by LR Test for Item Side Change Modeling Approach

Models	Preferred Model by LR test	by AIC	by BIC
No change (M1) vs Item-by-item (M4)	M4: χ^2 (df = 60) = 434.44, p < .001	M4	M4
No change (M1) vs Partial item-by-item (M5)	M5: χ^2 (df = 47) = 420.95, p < .001	M5	M5
No change (M1) vs Sub-content area (M3)	M3: χ^2 (df = 3) = 378.74, p < .001	M3	M3
No change (M1) vs Overall (M2)	M2: χ^2 (df = 1) = 367.91, p < .001	M2	M2
Partial item-by-item (M5) vs Item-by-item (M4)	M5: χ^2 (df = 13) = 13.49, p = .41	M5	M5
Overall (M2) vs Sub-content area (M3)	M3: χ^2 (df = 2) = 10.83, p < .01	M3	M2

Table 3. Results from the Interaction Model for Change

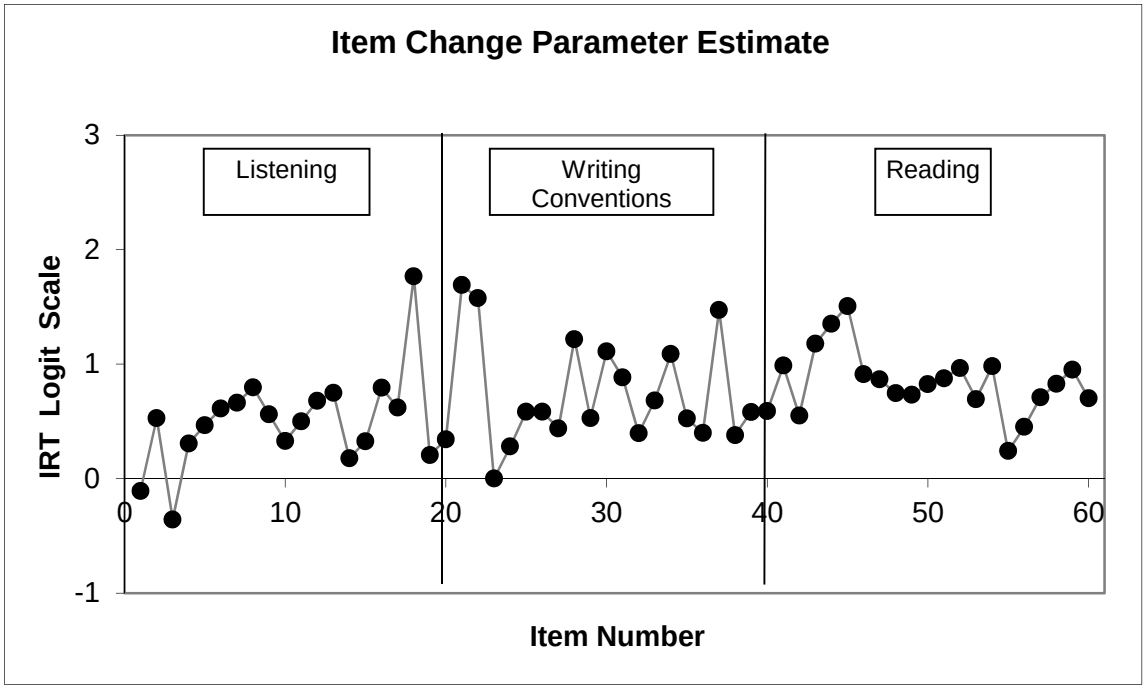
	Average Change	
	Estimate	Standard Error
Listening (L)	0.42	0.05
Writing Conventions (W)	0.72	0.05
Reading (R)	0.90	0.04

	Variances (diagonal) and Correlations (upper diagonal)			
	θ at Time Point 1	L	W	R
θ at Time Point 1	1.10	-0.60	-0.13	0.12
Listening (L)		0.34	0.40	0.29
Writing Conventions (W)			0.35	0.58
Reading (R)				0.39

	Model Fit			
Model	Deviance	No. of Par	AIC	BIC
Interaction Model for Change	22150.13	73	22296.13	22552.61

Figure 1

Item-by-Item Change Model: Item Change Parameter Estimate



Note. The black dot represents $\delta_{iC} = \delta_{iT1}$ (Year 2004) – δ_{iT2} (Year 2005).