Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy

Sunhee Kim[1*] and Gregory Camilli[2]

* Correspondence:
shk2125@columbia.edu
[1]College of Physicians and Surgeons, Columbia University, New York, NY, USA
Full list of author information is available at the end of the article

Abstract

Background: As the popularity of classroom observations has increased, they have been implemented in many longitudinal studies with large probability samples. Given the complexity of longitudinal measurements, there is a need for tools to investigate both growth and the properties of the measurement scale.

Methods: A practical IRT model with an embedded growth model is illustrated to examine the psychometric characteristics of classroom assessments for preschool children, and also to show how nonlinear learning over time can be investigated. This approach is applied to data collected for the Academic Rating Scale (ARS) in the literacy domain, which was administered on four occasions over two years.

Results: The model enabled an effective illustration of overall and individual gains over two academic years. In particular, a significant de-acceleration in latent literacy skills during summer was observed. The results also provided psychometric support for the argument that ARS literacy can be used to assess developmental skill levels consistent with theories of early literacy acquisition.

Conclusions: The proposed IRT approach provided growth parameters that are estimated directly, rather than obtaining these coefficients from estimated growth scores—which may result in biased and inconsistent estimates of growth parameters. The model is also capable of simultaneously representing parameters of items and persons.

Keywords: Item response theory; Latent growth models; Classroom assessment; Alternative assessment; Early literacy; Summer setback

In this paper, a hybrid statistical technique that embeds a growth model within a measurement model is introduced and applied. Specifically, we demonstrate that statistical estimates describing longitudinal growth can be obtained using a classroom-based set of items tapping cognitive skills while at the same time analyzing the psychometric properties of this instrument. As an illustration, we include an analysis of growth in language and literacy achievement for preschool children in a large-scale data set obtained as a national probability sample. Substantively, there is a literature on summer setback, but no studies we are aware of concerning this effect for preschool children. The purpose of this paper is to make the longitudinal modeling more accessible to researchers, but also to make a substantive contribution in an area having a sparse empirical literature on literacy development.

In previous research literature on developmental change, outcome measures are typically obtained as simple or weighted sums across the items in a particular assessment instrument. However, a simple aggregate score for investigating change has a number of potential problems as noted by Bereiter (1963) such as: paradoxical reliability of change scores; spurious negative correlation of change with initial status; and inconsistent scale units for change. These problems can be addressed partially with item response theory (IRT) modeling in which units are implicitly equated across measurement occasions, i.e., ability at different time points is transformed to a single scale (McArdle et al., 2009). This measurement model can be extended to investigate development or growth in a constant ability unit. We show that modeling repeated measurements in the framework of an integrated model also leads to a coherent set of tools for interpreting statistical estimates of growth. It will be demonstrated that substantive conclusions about development can be strengthened with information on measurement quality.

In contrast, simply collapsing item data into a single total score requires item analysis to be disjoint from the evaluation of growth. While one can obtain IRT scores independently of a growth model, those IRT *estimates* of ability must then be used to estimate growth parameters. One example of this is joint maximum likelihood estimation in IRT where estimates of ability are used to estimate item parameters (and vice versa). Unfortunately, item parameters are not consistent in this case (Embretson & Reise, 2000; Johnson, 2007), and this implies that growth parameters based on IRT estimates of ability may be biased even in the context of large data sets.

The measurement aspect of the model also provides tools to review the stability of each item over time. While many longitudinal studies in education provide classical measurement information on the quality of their measures (e.g., Cronbach's α), several studies (e.g., Embretson & Reise, 2000) have pointed out the limitations of classical test theory (CTT). For example, item difficulties are computed as the proportion correct and item discrimination is obtained as item-total correlations in CTT. These statistics are inconsistent for different samples with different ability levels, that is, these item parameters are not sample invariant. Integrated IRT models are designed to control for precisely this issue.

The purpose of this paper is to present the hybrid modeling approach and to demonstrate both the utility and limitations to developmental studies. After describing the model, an example will be presented in several stages. First, preliminary analyses are carried out to investigate model assumptions. Then different modeling approaches to change over time will be discussed, followed by substantive interpretations based on growth estimates.

Background

Measurement framework

Most IRT models contain two types of parameters: those for items, and those for person abilities. In the language of IRT, the probability of correct response on an item depends on person's ability and characteristics of an item such as difficulty. In this section, we introduce basic IRT models, and then hybrid IRT models for examining change over time. Embretson and Reise (2000), Hambleton, Swaminathan, and Rogers (1991), and Millsap (2010) provide more detailed information on IRT approaches.

1PL IRT model

In the IRT framework, Rasch models (Rasch, 1960), also known as one-parameter logistic (1PL) models, provide the probability of a correct answer of person i on item j (IRT models are easily adapted to polytomous assessment items as shown below). Let P_{ij} represent the true probability of a correct answer on a test question, the parameters of interest are the person's latent score θ_i and item difficulty b_j such that

$$P(Y_{ij} = 1|\eta_{ij}) = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}, \tag{1}$$

where $\eta_{ij} = \theta_i - b_j$.

The latent score θ_i represents individual ability, and is often scaled to be normally distributed with the mean zero and is typically specified with the condition $\theta \sim N(0,1)$. Item difficulty b_j indicates the level of θ having 50% chance to answer the specific item j correctly; thus, the higher (or lower) the item difficulty, the item requires higher (or lower) level of ability to get the item correct. Item difficulty in IRT shares the same scale as the person ability level, which allows the placement of item difficulties on the ability scale.

2PL IRT model

In addition to Rasch models, two-parameter (2PL) IRT models can be applied to longitudinal analysis. The 2PL model includes an additional item parameter, discrimination a_j. In the 1PL model, a constant discrimination a_j is assumed across items. If items vary with respect to discrimination, however, it may be useful to take this feature of item response function into account by substituting $\eta_{ij} = a_i(\theta_i - b_j)$ into Equation (1). Items with higher discriminations are more useful for separating examinees into different ability levels than items with lower discriminations, and can be useful for examining growth.

Modeling change

For the repeated administration of the same items to a sample over different occasions, Anderson (1985) introduced a multidimensional Rasch model with a time specific latent score θ_{it} for person i at occasion t, such that, a latent linear score can be defined as

$$\eta_{ijt} = \theta_{it} - b_j. \tag{2}$$

Anderson's approach is appropriate for understanding the impact of time on the ability distribution and for reviewing the characteristics of the test by time, Andrade and Tavares (2005) extended Anderson's model to 2PL and 3PL IRT models, which include item discrimination and item guessing parameters, respectively. However, Anderson's model does not explicitly contain change or growth parameters.

To reflect a person's differences in changes over T repeated test occasions, Embretson (1991) developed the multidimensional Rasch model for learning and change (MRMLC) which defines an individual propensity on occasion t proficiency θ_{it} as the sum of person's initial θ_{i1} and changes from ensuing test occasions of $\theta_{i2}, \ldots, \theta_{iT}$, such that,

$$\eta_{ijt} = \sum_{m=1}^{t} \theta_{im} - b_j, \tag{3}$$

with the IRT probability defined in Equation (1). While Andersen's (1985) model assumes the same items over occasions, Embretson's (1991) model allows some different items at different occasions to estimate changes in ability. Embretson (1991, 1997) also presented a 2PL version of the MRMLC that contains discrimination parameters.

The MRMLC estimates of individual changes are between one test occasion and the next, and thus the model is limited to investigating comparisons between two consecutive tests. In cases of the comparison between non-adjacent time points (e.g., the first test and the last test in four test occasions), a set of customized contrasts would be beneficial for examining specific substantive questions, which is the approach taken in the current paper. Specifically, functions of ability parameter are useful to understand changes over time, while the item parameters are assumed to be invariant (i.e., the performance of item parameters is identical over different measurement occasions).

IRT growth modeling

Children's developmental processes are stable for normal, healthy children, yet growth and development progress at different rates for *each* child (Gullo, 1994). In placing each student accurately on a developmental continuum, his or her trajectory in learning can be better understood and directed (Salinger, 2001). Estimates for *individual* learning and change (rather than changes in entire group) serve this purpose. The individual growth model (e.g., Verbeke & Molenberghs, 2000) is a relatively new statistical application in early childhood studies, and is pertinent to profiling the unique trajectories of individuals' abilities over time.

In this paper, a multidimensional IRT modeling approach (after Houts & Cai, 2013) is used to describe nonlinear and individual changes over time. The model can be specified in two steps. First, define the linear predictors for T assessment occasions for examinee i on item j as

$$\eta_{ijt} = a_j\left(\mu_t + \theta_{it} - \beta_j\right) \tag{4}$$

This indicates that the latent linear score for subject i on item j at time t is a function of item effects β_j, and person effects θ_{it} and μ_t. Here, a_j and β_j signifies the discrimination and difficulty of item j, θ_{it} represents a specific effect for individual i at time t that is a departure from average performance μ_t. Though as shown in Equation (4) the item difficulties β_j may vary, the growth modeling parameters μ_t and θ_{it} are constant for person i across items. This model is considered *multidimensional* because each occasion t provides a different measurement—even though the instrument remains constant. The model also allows different item sets for different time; which can be useful to investigate items that perform different for different time points or to utilize anchor items for different test occasions.

Second, an additional parameter of interest is the covariance structure of θ across different time points. There are various choices for specifying time dependencies in longitudinal IRT models (see Andrade & Tavares, 2005 for details). For example, an autoregressive model of order 1 has been employed in many time-series analysis, assumes that the correlations between the abilities decrease as the distances between the measurement occasions increase. While this model contains only two parameters, this

pattern of covariance has not been found in many longitudinal studies (Andrade & Tavares, 2005).

The current study employs an unstructured variance-covariance matrix except for the variance of ability at time 1, which is fixed at 1.0. One common way to fix the scale of θ is to dummy code the β parameters, where one item is omitted and serves as the reference item. In general, this strategy works only for 1PL models (e.g., Kamata, 2001). A more general method is to define the scale with the constraints $\mu_t = 0$ and $\sigma_t = 1$ for the first occasion ($t = 1$).

Non-linear changes

The form of the proposed model in Equation (4) can be usefully compressed for illustrating the role of contrasts to capture nonlinear change across time. Let the average proficiency on each occasion is $(\mu_1, \mu_2, \ldots, \mu_T)$. Given this representation, the model can be customized with time contrasts to aid interpretation effects holding substantive interest. To illustrate this concept, consider transformation, each defined by 4 coefficients across $T = 4$ measurement occasions, as:

$$\begin{aligned}
&\text{Intercept} \;: c_1 = \;\; 1, c_2 = \;\; 1, c_3 = \;\; 1, c_4 = \;\; 1\\
&\text{Contrast 1} : c_1 = \;\; 1, c_2 = -1, c_3 = \;\; 0, c_4 = \;\; 0\\
&\text{Contrast 2} : c_1 = \;\; 0, c_2 = \;\; 1, c_3 = -1, c_4 = \;\; 0\\
&\text{Contrast 3} : c_1 = \;\; 0, c_2 = \;\; 0, c_3 = \;\; 1, c_4 = -1,
\end{aligned}$$

with the constraint for any single contrast that

$$\sum_{i=1}^{4} c_i = 0. \tag{5}$$

Each of these contrasts represents the difference of two consecutive occasions (similar to Embretson's model). Once a set of contrast coefficients have been chosen for substantive reasons, custom effects can be calculated as

$$\psi = \sum_{i=1}^{T} c_i \mu_i. \tag{6}$$

These linear transformations can be written compactly in matrix form as

$$\mathbf{\Psi} = \mathbf{C}\boldsymbol{\mu}, \tag{7}$$

where C represents the transformation matrix above. Accordingly, the transformed covariance structure of $\theta(\mathbf{S}'_\theta)$ and parameter covariances $\left(\mathbf{S}'_p\right)$ can be obtained with

$$\mathbf{S}' = \mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-T}. \tag{8}$$

Standard errors are obtained as the square root of the diagonal elements of $\mathbf{S}'_p$. We recommend normalizing the rows of transformation matrix to preserve to the total variance across the separate measurement occasions.

Contrasts as illustrated above can also be applied to the individual proficiency estimates to examine individual growth trajectories consistent with substantive questions. While linear and quadratic time contrasts are widely used for longitudinal growth modeling, discontinuous patterns such as de-acceleration or setback, as seen during the summer vacation, have also been investigated in elementary and middle school settings

(Cooper et al., 1996; Entwistle & Alexander, 1992). The summer setback can be investigated by applying a customized contrast. By including categorical independent variables that reflect growth and changes in the model, our approach provides an appropriate tool for understanding change over time in school. In doing so, multiple indicators of growth (i.e., contrasts) can be obtained for further investigation of early learning processes.

The basic idea of this paper is familiar in the structural equation modeling (SEM) literature. For example, McArdle, Grimm, Hamagami, Bowles, and Meredith (2009) considered SEM models for representing change that include latent parameters for change in terms of mean and covariance structures as well as individual growth curves. Curran, Edwards, Wirth, Hussong, and Chassin (2007) also showed how item response theory can be incorporated into a modeling framework for analyzing growth. McArdle et al. (2009) conducted an analysis in which IRT was combined with latent growth curve analysis. A number of growth parameterization were formulated including no systematic change over time, linear change, and increasing-decreasing change over time. Both 2PL and partial credit models were discussed as possibilities for the measurement model, though a Rasch measurement model was used for their analysis. In this paper, we apply a number of these basic ideas in a slightly different way with more modern and efficient estimation techniques. While Curran et al. did consider item response theory (IRT) for analyzing growth, their approach was to estimate IRT scores and then apply growth modes to this data. In contrast, the approach taken below is based on an IRT model integrated into the growth model of this kind described earlier by Raudenbush and Sampson (1999), and MCArdle et al. (2009).

Estimation issues

This type of modeling has an advantage for many longitudinal data sets that equal numbers of measurements is not available for all subjects. In the case of incomplete or missing data, it can be shown that the correct likelihood and estimates for incomplete data can be obtained under the assumption of missing at random (Rubin, 1976). However, even with this advantage, the model specification on the covariance structure must be correct for estimating consistent parameters (Dmitrienko et al., 2005). Therefore, researchers need to select the model carefully. One method is to check the model-fit statistics, e.g., AIC, and BIC, to find the best fit model.

Several researchers have employed IRT models embedded within a larger modeling framework. Previously, the Rasch model had been proposed for specifying multilevel models suitable for longitudinal analysis (e.g., Briggs & Wilson, 2007; Kamata, 2001; McArdle et. al, 2009; Meiser, 1996; Sampson & Raudenbush, 2004). In the recent literature, more complex models are considered that contain more parameters (for both items and growth). For example, von Davier, Xu, and Carstensen (2011) introduced a general model framework and a growth-mixture MIRT model that allows for variation in latent trajectories across clusters in a nested sample.

Both 1PL and 2PL models of these kinds can be implemented in generalized linear mixed or mixture modeling software with EM or maximum likelihood estimation (e.g., Zheng & Rabe-Hesketh, 2007; Vermunt & Magidson, 2008; von Davier, Xu & Carstensen, 2011). While EM estimation is generally preferred to maximum likelihood, EM code requires lengthy execution time for models with higher dimensionality. In this paper, we

demonstrate the software package flexMIRT employing the Metropolis Hastings Robbins Munro (MHRM) algorithm which is highly efficient in terms of computation time and is easily implemented (Houts & Cai, 2013). However, model fit issues are still under investigation with respect to the MHRM algorithm, and item fit statistics are not yet provided by flexMIRT (Houts & Cai, 2013. p. 89) Additional file 1.

Methods

To illustrate the application of an IRT growth model, a real data example is provided below concerning an important question in literacy/language development. A number of educators have urged the use of classroom assessments, which are measurements designed to capture learning process are necessary for accurately describing learning in school (e.g., NAEYC, 1992; Schweinhart, 2003; Scott-Little, Kagan & Frelow, 2003; Shepard, Kagan & Wurtz, 1998). Although these types of assessment are currently being used by a number of school systems and are increasingly becoming a part of state assessment packages (Jones, 2003; NRC, 2008), none that to our knowledge have been psychometrically evaluated with regard to the developmental characteristics of children. In addition, previous studies on changes during the summer (Burkam et al., 2004; Rock, Pollack & Weiss, 2004) of kindergarten to first-grade growth have employed scores of a standardized cognitive test not a classroom assessment. The model proposed herein provides estimates of both item and growth parameters as well as providing tools for the evaluation of item quality of a classroom assessment.

In the following sections, information is provided on the data source, sample, measurement instrument, and formal statistical model. This is followed by an elaboration of the IRT models for polytomous items, investigation of measurement properties of an instrument for assessing growth, examining of growth modeling results, and finally substantive discussion of findings and implications.

Data source

The public-use version of the data from the National Center for Early Development and Learning Multi-State Study of Pre-Kindergarten 2001–2003 (Clifford et al., 2009) is used. The original study was designed to examine the effects of variations in pre-kindergarten and kindergarten experiences on children's social and academic outcomes.

Sample

In the given data, children were randomly selected, four per classroom, from 40 classrooms in each of six states. The classroom samples were stratified within each state according to programs in schools versus another site; full-day versus part-day program; and teachers with and without a college degree. The children were followed from the beginning of pre-kindergarten (P) through the end of kindergarten (K). Complete data were collected from 778 (81%) children, partial data were collected from 132 (14%) children, and no data were obtained from 49 (5%) of the children. All of the sample cases in the data are included in the analysis. Sampling weight provided with the data set is used in the present study. The original sampling weights were rescaled by dividing the product of each weight and the sample size over the summation of all the weights (i.e., the population size). Normalizing the sample weights (so they add up to the actual sample size) is a generally recommended procedure (Pfeffermann et. al, 1998).

Measurement instrument

On four occasions, in the fall and spring of preschool and kindergarten, each child was assessed by the Academic Rating Scale (ARS). The National Center for Education Statistics (NCES, 2000) developed the ARS for teachers' evaluations of young students' academic achievement in three domains: language/literacy, general knowledge, and mathematical thinking. The Rasch person reliability of the ARS in kindergarten was reported with a range of .87 to .91 (NCES, 2002). The ARS includes items designed to measure both the process and products of children's learning in school, therefore it represents broader lends on school outcomes than standardized achievement batteries (NCES, 2002). Because the assessment of early literacy skills is a very controversial topic in early-childhood education (IRA & NAEYC, 1998; Neuman & Dickinson, 2001; Salinger, 2001), the language/literacy domain is analyzed in this paper. The domain consists of nine items on a 5-point Likert scale (i.e., the items are 'polytomous' in psychometric terms, such that, 1: Not yet, 2: Beginning, 3: In Progress, 4: Intermediate, 5: Proficient). The mean score (based on the total-sum) at the four assessments are shown in Table 1 and the individual item descriptions are presented in Figure 1 as a part of the item map.

Item response model

For the analysis, Samejima's graded response model (Thissen & Steinberg, 1986) was used for all items. In the model, the probability of obtaining score k for an item with k ordinal response categories for person i on item j at time t can be specified as

$$P(Y_{ijt} = k|\eta_{ijt}) = P(Y_{ijt} \le k|\eta_{ijt}) - P(Y_{ijt} \le k-1|\eta_{ijt}). \quad (9)$$

Typically, a set of $k - 1$ category intervals are estimated as b_k for each item j, namely item location parameters. The a parameters may be constrained to be equal or vary across items. Since the measure is used observationally by teachers, the guessing parameter (in 3 PL IRT model) is not considered in the analysis.

Differential item functioning (DIF) over test occasions

We sought to establish that item parameters were invariant across the two years of this study. In other words, one needs to establish the invariance assumption of item parameters across pre-kindergarten through kindergarten to apply IRT. Differential item functioning (DIF) techniques can be used to detect items that behave differently across test occasions. More general reviews on DIF can be found in Camilli and Shepard (1994) and French and Miller (1996). Preliminary to DIF analysis, both 1PL and 2PL reference models were obtained. The results are shown in the top half of Table 2. For both AIC and BIC, the 1PL model fit better, and for this reason it was decided to use the 1PL model for investigation of DIF.

Table 1 Summary of ARS mean scores

Time	$\bar{X}$	*SD*	*n*
P Fall	2.09	0.81	916
P Spring	2.79	0.92	941
K Fall	2.40	0.92	827
K Spring	3.67	0.95	817

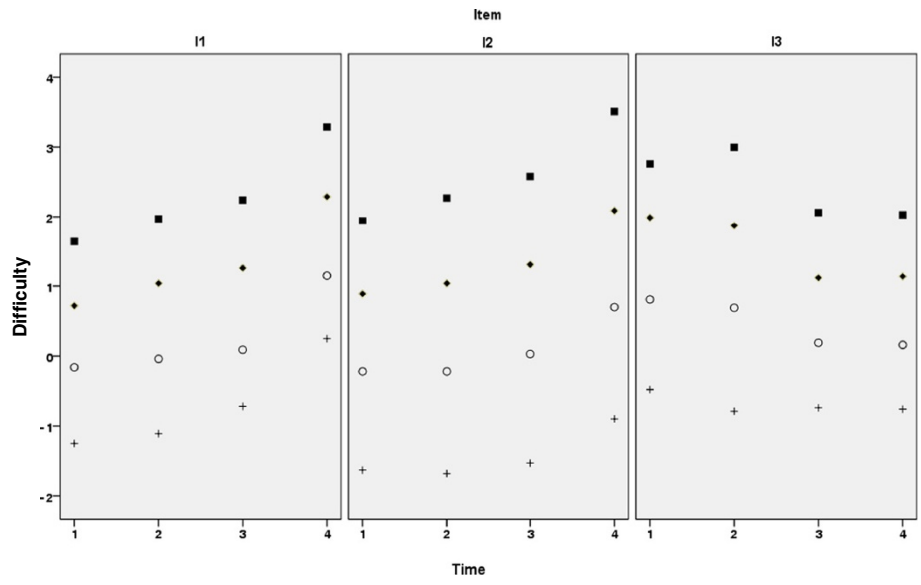

Figure 1 Difficulty drift for items 1–3.

To screen the data for potential DIF by time, we allowed the difficulty locations for each item (holding other items fixed) to vary by occasions. In Table 3, AIC and BIC fit values (Akaike and Bayesian information criterion, where smaller values indicate better fit) are reported by item, along with the change from the base model in which all item difficulties are constrained to be equal across occasions. The first three items have the largest DIF indices (i.e., they increase fit the most), and this was corroborated by applying other DIF approaches. To account for this DIF, we report two versions of each analysis for the 1PL and 2PL models: one in which all item difficulties were constrained to be equal across occasions, and one in which items 1–3 were allowed to fit differently across occasions. The DIF is illustrated in Figure 2. It can be seen that for items 1 (Uses complex sentence structures) and 2 (Understands stories/texts read to her/him), all difficulty locations drifted up over time, while the locations for item 3 (Easily names all upper/lower alphabet letters) drifted down.

Modeling growth

To analyze longitudinal changes in latent skills as well as the psychometric characteristics of ordinal items that measures process in classroom, we employed a model combining an IRT model for polytomous items with a model for change over time.

Table 2 Model fit statistics

		Model Fit	
DIF	**Model**	**AIC**	**BIC**
No DIF Adjustment	1PL	69456.0	69696.9
	2PL	70302.8	70582.9
DIF Adjusted	1PL	68018.8	68606.5
	2PL	69209.6	69666.7

Note. A smaller AIC or BIC indicates a better-fitting model.

Table 3 DIF statistics (1PL model)

	Item	AIC	BIC	Change
All Constrained		69456.0	69696.9	–
Free	1	68898.0	69197.8	−558.1
	2	69120.3	69420.1	−335.7
	3	69057.2	69357.0	−398.8
	4	69286.3	69586.1	−169.7
	5	69438.2	69738.0	−17.9
	6	69430.4	69730.2	−25.7
	7	69301.7	69601.5	−154.3
	8	69273.8	69573.6	−182.3
	9	69415.3	69715.1	−40.8
	1-3	68179.6	68597.3	−1276.5

Note. Change statistics are based on the simulated distribution of −2 log likelihood of the fitted model. The change values in the last column can be roughly interpreted as chi-square statistics with 12 degrees of freedom.

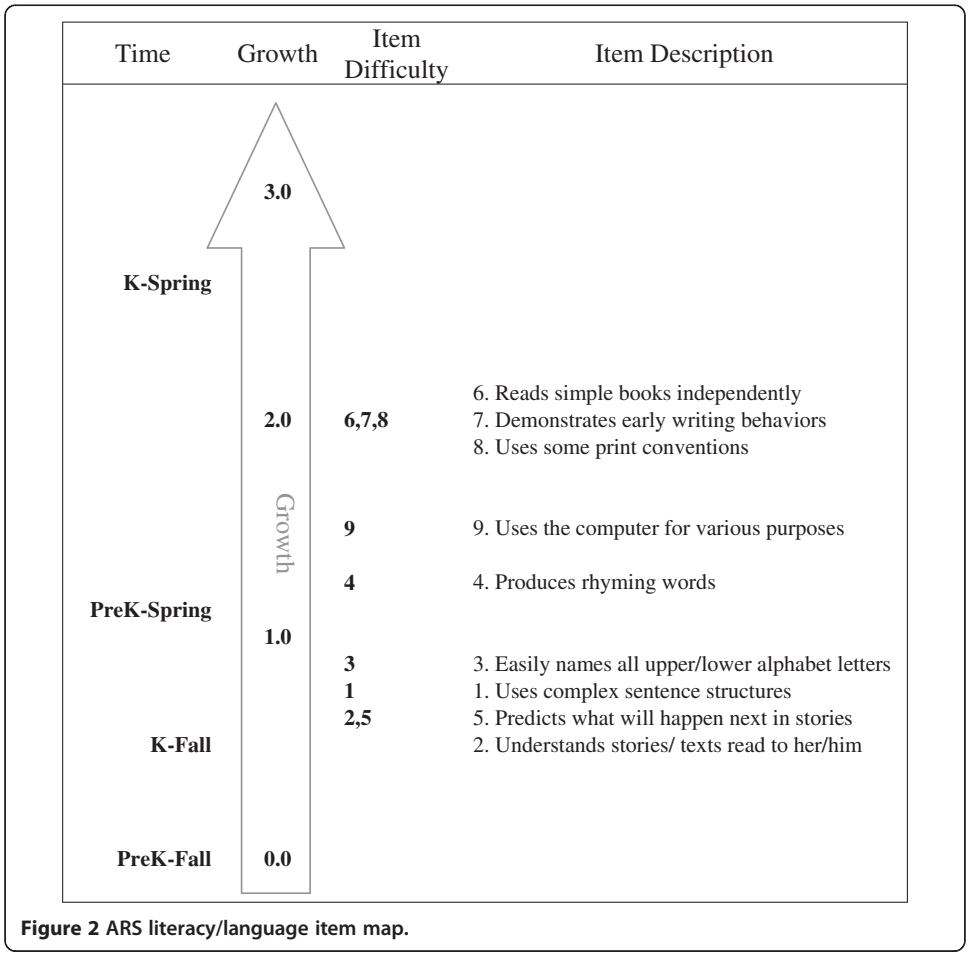

Figure 2 ARS literacy/language item map.

Growth parameters

The parameters of substantive interest are the person parameters and their covariance across four occasions. Customized contrasts were used to define parameters in a way that directly addressed the issue of summer setback. For the first set of contrasts, we designated performance in the fall of pre-school (P Fall), the first test occasion, as a reference. Thus, growth is estimated relative to the first test occasion. Accordingly, the proficiency level of a specific person is also estimated relative to this reference. This set of contrast codes is given in the top panel of Table 4.

For the second set of contrasts, a set of three orthogonal contrasts (C_2 through C_4) were chosen to represent growth and individual effects across measurement occasions in addition to an average ability obtained by C_1, as displayed in the bottom panel of Table 4. The summer effect contrasts children's scores on the second and third testing occasions, which represent the change during the summer prior to kindergarten, as shown in contrast C_2. The differences in the rates of growth from the first year and the second year are estimated by the last contrast C_3. Finally, the overall change C_4, is estimated by the difference between P Fall and K Spring performances.

Design

Three major facets are examined to enhance the validity of the results. First, estimates were obtained with both the 1PL and 2PL models; Second, two types of contrast are reported; and third, both of these conditions are crossed with the original condition and no DIF conditions in which three items were allowed to fit differently across occasions.

Results

In this section, we first provide the results from the growth component of the model. Then the measurement components of the model are presented, along with description of how these results can be used as validity evidence for the ARS scale. Analog to IRT model parameter specifications, the person parameters are used to review the growth or changes over time, the item parameters are used to review the item characteristics, and then the combinations of those two types of parameters are used to review time-appropriate use of items.

Growth results

All growth parameters within the T and C contrast types were similar as shown in Table 5, except for the 2PL model with DIF. With DIF adjustment, the 2PL model

Table 4 Contrast codes for models

T Contrasts	T_1	T_2	T_3	T_4
P Fall	1	0	0	0
P Spring	0	1	0	0
K Fall	0	0	1	0
K Spring	0	0	0	1
C Contrasts	C_1	C_2	C_3	C_4
P Fall	1/2	0	1/2	$-1/\sqrt{2}$
P Spring	1/2	$-1/\sqrt{2}$	−1/2	0
K Fall	1/2	$1/\sqrt{2}$	−1/2	0
K Spring	1/2	0	1/2	$1/\sqrt{2}$

Note. The contrasts in Model 2 are labeled as: C_1 is the average ability; C_2 is the summer effect; C_3 is the overall change; and C_4 represents the increment in growth rate from preschool to kindergarten.

Table 5 Growth parameter estimates

			T Contrast		C Contrast	
DIF	**Model**	**Effect**	**Estimate**	**SE**	**Estimate**	**SE**
No DIF Adjustment	1PL	2	1.09	.03	-.48	.04
		3	.41	.04	.38	.08
		4	2.26	.03	1.60	.04
	2PL	2	1.02	.03	-.39	.04
		3	.47	.04	.37	.08
		4	2.22	.04	1.57	.06
DIF Adjusted	1PL	2	1.16	.04	-.50	.06
		3	.45	.04	.50	.08
		4	2.60	.06	1.84	.08
	2PL	2	.58	.04	-.33	.06
		3	.12	.03	.37	.06
		4	1.43	.07	1.01	.10

Note. All coefficients are significant at $\alpha = .001$.

obtained substantially different estimates of both the summer effect and overall growth. For the 1PL model, growth parameters under the DIF and no DIF conditions were similar. Given that the 1PL model with DIF adjustment showed the best fit, this condition is used to examine growth below.

Accordingly, growth begins a zero baseline in preschool fall (PF), increases to 1.16 in preschool spring (PS), decreases to .45 in kindergarten fall (KF), and increases to 2.6 in kindergarten spring (KS) on average. The variation in proficiency as shown in Table 6 increases substantially after KF. It can also be seen in the covariance matrix in Table 6 that proficiency within school year is positively correlated, but much higher in kindergarten than preschool.

From the customized contrasts, the summer effect is −.5, that is, the estimate corresponding to C_2. Rescaled to the standard deviation at PS, the effect size is $d = -.5/\sqrt{1.2} = -.46$. This indicates that the average ARS literacy/language ability diminishes substantially over the summer. However, the overall change in average child proficiency level is $d = 1.85$ from PF to KS is relatively large, indicating that despite the summer loss, children tend to develop rapidly over the course of kindergarten. The variability of C_2 can be illustrated in personal profiles over time, as shown in Figure 3. The proficiency estimates of one child (ID3) declined over the summer, while those of other children were improved with

Table 6 1PL covariance matrices (DIF adjustment)

Model 1	T_1	T_2	T_3	T_4
T_1	1.00			
T_2	.71	1.66		
T_3	.47	.59	1.91	
T_4	.35	.46	1.30	1.66
Model 2	C_1	C_2	C_3	C_4
C_1	3.50			
C_2	.30	1.2		
C_3	-.35	.12	.56	
C_4	.44	.54	.02	.98

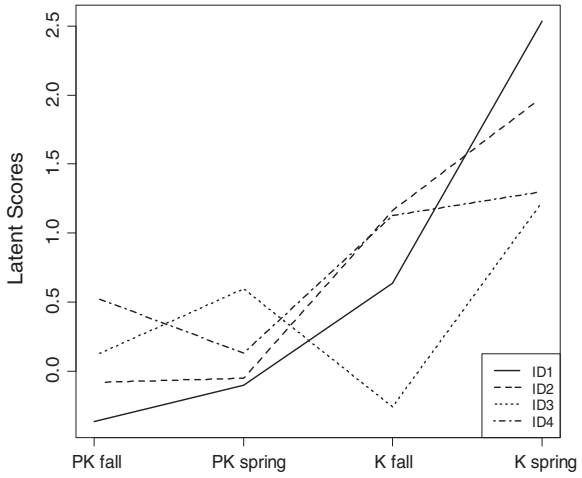

Figure 3 Examples of person changes over time incorporating fixed and random effects.

different rates. Those individual scores can be obtained directly by signifying the appropriate option in flexMIRT.

Another effect if interest is C3, which indicates differential growth across preschool and kindergarten. The summer effect is only slightly correlated with differential growth (see the covariance between C_2 and C_3 in Table 4). This indicates that a student who learns more during the summer does not necessarily have growth rate during the subsequent academic year. However, students who retain more skills (as measures by the ARS) over the summer tend to have higher overall growth ($r = .49$). This result can be seen in Figure 3, by comparing the profiles for ID1, ID2, and ID4 with that of ID3.

Measurement results

Item parameters

Estimated location parameters are given in Table 7 for the 1PL DIF model.

The relative difficulties among items can be examined in terms of the average item location parameters. Similar to item difficulty parameters, high values of item location estimates indicate that the item is harder than others. Though the average b parameters are slightly higher than those for typical tests, it should be realized that the test scale was standardized to P Fall. The lower item locations fall comfortably within the range preschool proficiency as shown in Table 7.

Table 7 1PL item parameter estimates with DIF adjustment (a = 1.7)

	Average difficulty	Locations			
Item	b	b_4	b_3	b_2	b_1
1	0.79				
2	0.64		see Figure 2		
3	0.94				
4	1.42	−0.35	0.86	2.06	3.11
5	0.71	−1.48	0.10	1.39	2.83
6	1.88	0.18	1.37	2.47	3.49
7	1.92	0.25	1.44	2.50	3.49
8	1.94	0.18	1.39	2.52	3.67
9	1.50	−0.65	0.82	2.28	3.54

Characteristics of the ARS

Based on the model outcomes, we review the developmental characteristics of ARS. First, the results of the current study support the claim that the ARS is measuring a developmental proficiency. The developmental trend is similarly demonstrated in several select individual proficiency profiles in Figure 3.

Second, the estimated magnitudes of the item difficulties support the theoretical order of emergent literacy (McGee & Richgels, 2004; Sutzby, 1985). We present an item map in Figure 1, which includes the average location parameters in Table 7. Items relevant to features of novice readers/writers (items 1, 2, 5) had lower item locations than those of experienced readers/writers (items 6, 7 and 8).

Third, our results indicate that the ARS items are properly designed for assessing target graders. Although ARS has been used for studying children's growth in Pre-K through kindergarten, NCES originally designed it to measure grade K students' literacy and language skills. As shown in the item map, all item locations are within the growth interval from fall to spring of kindergarten. Meanwhile, it is evident that many items are moderately to very difficult in the fall of preschools, thus limiting their capacity to assess literacy proficiency. The estimate of positive difference within years also supports this argument. The outcome indicates higher within year growth in kindergarten than those in preschool; in other words, on average children in kindergarten gain more than children in preschool.

Conclusions

In this paper, we presented an application of IRT growth modeling to illustrate nonlinear individual learning and change over time and to investigate the psychometric properties of a classroom assessment. A particularly striking advantage of the proposed IRT approach is that growth parameters can be estimated directly, rather than obtaining these coefficients from estimates of ability. Estimates of growth based on estimates of ability may build systematic measurement error into the former. The model is also capable of simultaneously representing parameters of items and persons, including their individual changes, in embedded IRT models.

By integrating measurement and growth, the proposed approach provides more accurate estimates for overall and individual changes as well as the variation of growth. For instance, we identified several items that perform differently across assessment occasions. The inclusion of those items had a notable affect on growth estimates when DIF was implemented in the 2PL model. At the same time, our approach provides a method to investigate the item characteristics controlled for nonlinear growth. The validating developmental features of ARS are investigated through IRT modeling in the current paper, which does not require either linear or even monotonic growth. Still, the significant improvement of personal trait levels over years indicates that the ARS literacy domain can be used to study the progress of emerging readers during preschool and kindergarten. Children in kindergarten grew more and faster, and the test was targeted to kindergarten, this implies that the ARS is functioning according to expectation. Some support is thereby offered that items are indeed more representative skills of the kindergarten level language/literacy domain. Also, the items can document early literacy development along a continuum that has a thread of connection with early literacy acquisition theories. The item map provides educators useful information about the literacy skill difficulty level, which can be applied to their educational practices and curriculum.

The integrated approach is also useful to profile a person's skill level at each repeated test occasion similar to other latent growth IRT model in previous studies (e.g., Curran et al., 2007). We demonstrated a personal profiling (as shown in Figure 3) with the personal changes over time to review the trend of changes over time. Personal profiling is a method of monitoring individual changes especially with missing data. Further, individual estimates based on a non-linear growth model may provide a more accurate representation than parametric approaches (e.g., linear and quadratic trends). For instance, with assuming a linear growth, the setback during the summer is misrepresented in the growth estimates. Time contrasts can be customized for certain hypotheses or assumptions about growth. Although parallel effects can be tested with raw scores, these approaches have limitations in the longitudinal inquiry (Bereiter, 1963). In particular, scores would need to be equated to the same scale over different time points for longitudinal analysis.

Both the 1PL and 2PL models can be implemented practically with the MHRM algorithm, though flexMIRT is new and not yet highly accessible to many researchers focusing on the school settings. Only a handful of quadrature points could be used for EM estimation (a maximum of 13 on a computing platform with 8GB of memory using 8 threads). Even then, convergence required about two hours whereas the same problem required less than 2 two minutes with the MHRM algorithm. Thus, lack of familiarity with new estimation software will clearly be overcome by practical utility.

The current study utilized the sampling weight provided in the original data set. However, the nested sampling structure was not fully reflected in our analysis, partially because the class identifier was not available. Also, full information from the sampling design would be necessary to compute standard errors with precision (which could conceivably be underestimated by a factor of 2 through 3; Johnson & Rust, 1992). Although the software does not provide a mixture modeling, the idea of utilizing customized contrasts can be extended to the previous literature considering mixture distributions of growth (e.g., von Davier, Xu & Carstensen, 2011). Another extension of this model might include the consideration of rater effects, especially with constructed response items (e.g., ARS items) for the future studies.

Summer setback after kindergarten has been observed in previous studies (Burkam et al., 2004; Rock, Pollack & Weiss, 2004) but summer setback prior to kindergarten has not, to our knowledge, been investigated previously. One of the interesting findings of this paper is that literacy/language proficiency for preschoolers diminishes substantially during the summer before kindergarten—even though positive overall growth is demonstrated in preschool and kindergarten. Another important result regarding to the summer effect is that summer setback effect is much larger for some children than others. Some perspective is required to understand the practical significance of this finding. From Table 2, it can be seen that the summer setback effect (–.50 units) is about 27% of the overall change (1.84 units) from P Fall to K Spring. However, there is substantial individual variation in the setback effect. Using a normal distribution as a rough guide, about 9% of preschoolers have a setback effect of one standard deviation or more in terms of overall ability as represented by C1, which tends to place them at more risk of performing poorly in kindergarten. If these students would simply maintain reading their level of achievement in literacy/language over the summer, a substantial, if not dramatic, improvement might be obtained. However, the story is more

complex. Average growth from fall to spring in kindergarten is almost one standard deviation. Thus, many children appear to cover lost ground rapidly, and the summer setback effect moderately correlated with overall growth. While some children grow much faster than other in kindergarten than in preschool, this effect is not highly correlated with other growth measures. The current study does not shed much light on differential growth. Summer academic support is an important problem, and targeting support for summer growth is a key issue. The relatively large variability of the summer effect implies that further research on the source of variation is urgently needed. Previous literature (Burkam et al., 2004; Rock, Pollack & Weiss, 2004) reported that disadvantaged kindergarten children tend to lose more on scores during the summer vacation, and this suggest the estimate of the summer effect could be used to explore background variables that indicate access of children to academic support.

Additional file

Additional file 1: Example flexMIRT code.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SK and GC reviewed the literature, designed and carried the analyses, and prepared the manuscript. All authors read and approved the final manuscript.

Author details

[1]College of Physicians and Surgeons, Columbia University, New York, NY, USA. [2]Graduate School of Education, Rutgers University, New Brunswick, NJ, USA.

Received: 3 September 2013 Accepted: 22 November 2013
Published: 14 January 2014

References

Andersen, EB. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16.

Andrade, DF, & Tavares, HR. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis, 95*, 1–22.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In CW Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Briggs, DC, & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*(2), 131–155.

Burkam, DT, Ready, DD, Lee, VE, & LoGerfo, L. (2004). Social-class differences in summer learning between kindergarten and first grade: model specification and estimation. *Sociology of Education, 77*(1), 1–31.

Camilli, G, & Shepard, LA. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Clifford, RM, Burchinal, M, Howes, C, Winton, PJ, Bryant, DM, Barbarin, O, & Early, DM. (2009). *National center for early development and learning multistate study of pre-kindergarten, 2001–2003 [computer file]. ICPSR04283-v2*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. 10.3886/ICPSR04283.v2.

Cooper, H, Nye, B, Charlton, K, Lindsay, J, & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227–268.

Curran, PJ, Edwards, MC, Wirth, RJ, Hussong, AM, & Chassin, L. (2007). The incorporation of categorical measurement models in the analysis of individual growth. In T Little, J Bovaird, & N Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies of human development* (pp. 89–120). Mahwah, NJ: LEA.

Dmitrienko, A, Molenberghs, G, Chuang-Stein, C, & Offen, W. (2005). *Analysis of clinical trials using SAS: a practical guide*. Cary, NC: SAS Institute Inc.

Embretson, SE. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495–515.

Embretson, SE. (1997). Structured ability models in tests designed from cognitive theory. In M Wilson, G Engelhard Jr, & K Draney (Eds.), *Objective measurement: theory into practice* (Vol. 4, pp. 223–236). Greenwich: Ablex.

Embretson, SE, & Reise, SP. (2000). *Item response theory for psychologists*. NJ: Lawrence Erlbaum Associates Inc. Publishers.

Entwistle, DR, & Alexander, KL. (1992). Summer setback: race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review, 57*, 72–84.

French, AW, & Miller, TR. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315–332.

Gullo, DF. (1994). *Understanding assessment and evaluation in early childhood education*. New York: Teachers College Press.

Hambleton, RK, Swaminathan, H, & Rogers, HJ. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Houts, CR, & Cai, L. (2013). *flexMIRT user's manual version 2.0: flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

International Reading Association (IRA) and NAEYC. (1998). *Joint position statement on learning to read and write: developmentally appropriate practices for young children*. Washington, DC: NAEYC.

Johnson, MS. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, 20*(10), 1–24.

Johnson, EG, & Rust, KF. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190.

Jones, J. (2003). *Early literacy assessment systems: essential elements*. Princeton, NJ: Educational Testing Service.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

McArdle, JJ, Grimm, K, Hamagami, F, Bowles, R, & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*, 126–149.

McGee, LM, & Richgels, DJ. (2004). *Literacy's Beginnings: supporting young readers and writers* (4th ed.). Needham, MA: Allyn and Bacon.

Meiser, T. (1996). Logninear rasch models for the analysis of stability and change. *Psychometrica, 61*(4), 629–645.

Millsap, RE. (2010). Testing measurement invariance using IRT in longitudinal data: an introduction. *Child Development Perspectives, 4*(1), 5–9.

NAEYC & NAECS/SDE (National Association of Early Childhood Specialists in State Departments of Education). (1992). Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. In S Bredekamp & T Rosegrant (Eds.), *Reaching potentials: appropriate curriculum and assessment for young children* (volume 1st ed., pp. 9–27). Washington, DC: NAEYC.

National Center for Education Statistics. (2000). *Early childhood longitudinal study, kindergarten class of 1998–99, data files and electronic code book, ECLS-K base year public-use (NCES 2001–029) 2002134 [CD-ROM]*. Washington, DC: Author.

National Center for Education Statistics. (2002). *ECLS-K first grade public-use data files and electronic code book (NCES 2002-134) [CD-ROM]*. Washington, DC: Author.

National Research Council. (2008). *Early childhood assessment: why, what, and how*. Committee on developmental outcomes and assessments for young children. In CE Snow & SB Van Hemel (Eds.), *Board on children, youth, and families, board on testing and assessment, division of behavioral and social sciences and education*. Washington, DC: The National Academies Press.

Neuman, SB, & Dickinson, DK. (2001). *Handbook of early literacy research*. New York: Guilford.

Pfeffermann, D, Skinner, CJ, Holmes, DJ, Goldstein, H, & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*, 23–40.

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

Raudenbush, SW, & Sampson, R. (1999). Assessing direct and indirect associations in multilevel designs with latent variables. *Sociological Methods & Research, 28*(2), 123–153.

Rock, DA, Pollack, M, & Weiss, M. (2004). *Assessing cognitive achievement growth during the kindergarten and first grade years (ETS RR-04-22)*. Princeton, NJ: ETS.

Rubin, DB. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Salinger, T. (2001). Assessing the literacy of young children: the case for multiple forms of evidence. In S Neuman & D Dickinson (Eds.), *Handbook of early literacy research* (pp. 421–443). New York, NY: Guilford Press.

Sampson, RJ, & Raudenbush, SW. (2004). The social structure of seeing disorder. *Social Psychology Quarterly, 67*(4), 319–342.

Schweinhart, LJ. (2003). Issues in implementing a state preschool program evaluation in Michigan. In C Scott-Little, SL Kagan, & RM Clifford (Eds.), *Assessing the state of state assessments: perspectives on assessing young children* (pp. 37–42). Greensboro, NC: SERVE.

Scott-Little, C, Kagan, SL, & Frelow, V. (2003). *Standards for preschool children's learning and development: who has standards, how were they developed, and how are they used?* Greensboro, NC: SERVE.

Shepard, LA, Kagan, SL, & Wurtz, E (Eds.). (1998). *Principles and recommendations for early childhood assessments*. Washington, DC: National Goals Panel.

Sulzby, E. (1985). Kindergartners as writers and readers. In M Farr (Ed.), *Advances in writing research: volume 1.children's Early writing development* (pp. 127–199). Norwood, NJ: Ablex.

Thissen, D, & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–576.

Verbeke, G, & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Vermunt, JK, & Magidson, J. (2008). *LG-syntax user's guide: manual for latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.

Von Davier, M, Xu, X, & Cartensen, CH. (2011). Measuring growth in a longitudinal large scale assessment with a general latent variable model. *Psychometrika, 76*(2), 318–336.

Zheng, X, & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models using gllamm. *The Stata Journal, 7*, 313–333.

doi:10.1186/2196-0739-2-1
Cite this article as: Kim and Camilli: **An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy.** *Large-scale Assessments in Education* 2014 **2**:1.