Using POMDPs to Forecast Kindergarten Students Reading Comprehension

Russell G. Almond
Educational Psychology and Lerning Systems
Flordia State University
Tallahassee, FL 32306
ralmond@fsu.edu

Umit Tokac
Educational Psychology and Lerning Systems
Flordia State University
Tallahassee, FL 32306
ut08@my.fsu.edu

Stephanie Al Otaiba*
Department of Teaching and Learning
Southern Methodist University
Dallas, TX 75275
salotaiba@smu.edu

Abstract

Summative assessment of student abilities typically comes at the end of the instructional period, too late for educators to use the information for planning instruction. This paper explores the possibility of using Hierarchical Linear Models to forecast students end of year performance. Because these models are closely related to partially observed Markov decision processes (POMDPs), these should support extensions to instructional planning to meet educational goals. Despite the new notation, the POMDP models are subject to a familiar problem from the educational context: scale identifiability. This paper describes how this problem manifests itself and looks at one potential solution.

1 INTRODUCTION

There is a long tradition in education of separating instruction and assessment: summative assessment of what a student learns comes at the end of the unit/semester/year. As limited time is allocated for assessment, such assessments are typically limited in their reliability (accuracy of measurement) and content validity (coverage of the targeted knowledge, skills and ability). Because summative assessment comes at the end of instructions, instructors are not able to make changes to their instructions to maximize student learning (Almond, 2010).

Bennett (2007) suggested breaking the summative assessment into four or six periodic assessments. First, spreading the cost (student time taken away from direct instruction) over multiple measurement occasions allows for longer testing providing both greater content coverage and reliability. Moreover, a proper model for student growth allows forecasting of the students eventual status at the end of the year. Consequently, teachers and administrators can form plans for students which maximize learning outcomes and identify students for whom the goals are unreachable for special instruction. In this sense, the periodic assessments play a role somewhere between traditional summative assessment and formative assessment — assessing student learning for the purpose of improving instruction (Black & Wiliam, 1998; Wiggins, 1998; Pelligrino, Glaser, & Chudowsky, 2001).

Almond (2007) noted that the forecasting could be done using a partially observed Markov decision process (POMDP; Boutilier, Dean, & Hanks, 1999): the latent variables describing student proficiency form an unobserved Markov process, and the periodic assessments provide observable evidence about the state of those latent variables. The instructional activities chosen between time points are the measurement space, and in fact, the students response to instruction often provides important clues about their proficiency and specific learning problems (Marcotte & Hintze, 2009). Almond (2009) notes the similarity between POMDPs and other frameworks more commonly used in education, such as latent growth modeling (Singer & Willett, 2003) and hierarchical linear modeling (HLM; Raudenbush & Byrk, 2002). The principle difference is one of emphasis: in the POMDP framework, the emphasis is usually on estimating the individuals latent state for the purpose of planning. In the HLM and multilevel growth model, the emphasis is usually on estimating the effectiveness of various activities. This paper looks at the problem of forecasting using HLM models both directly and through conversion to POMDP parameterizations.

The purpose of our study is to try to fit a POMDP-based latent growth model using Bayesian methods to a set of data documenting the development of Reading skills in a number of Kindergarten students. Once the

*Some of the work took place while she was at the Florida Center for Reading Research, Tallahassee, FL

model is successfully fit, we will use it to predict the end-of-year status of the students.

2 THE DATA

This study uses longitudinal data about reading development originally collected by the Florida Center for Reading Research (Al Otaiba et al., 2011). The reading skills for this initial cohort of students was measured three times (Fall, Winter and Spring) during Kindergarten, and follow-up measurements were taken at the end of 1st, 2nd and 3rd grade. There were 247 students in the initial sample, but only 224 were still in the area at the end of the first year.

During Kindergarten, children rapidly develop in Reading and pre-Reading skills (e.g., oral vocabulary and letter identification). Consequently, not all measures are appropriate for all time points. Consequently, different measures were collected at different time points. Table 1 shows the measures that were collected during Kindergarten:

Table 1: Measures Collected By Occasion

Measure	Fall	Winter	Spring
LW	X	X	X
PV	X	X	X
ISF	X	X	
PSF		X	X
NWF		X	X
LNF	X	X	X

The measures are taken from the Woodcock-Johnson III Cognitive Test (WJ-III; Woodcock, McGrew, & Mather, 2001) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminkski, 2002). The measures used were:

LW – Letter-Word Identification (WJ-III)

PV – Picture Vocabulary (WJ-III)

LNF – Letter Naming Fluency (DIBELS)

ISF – Initial Sound Fluency (DIBELS)

PSF – Phoneme Segmentation Fluency (DIBELS)

NWF – Nonsense Word Fluency (DIBELS)

The Woodcock-Johnson measures are available in several forms. We used the "W" scale (which is scaled to an item response theory model), as it showed more variation than the scale scores.

Additionally, teacher and school identifiers are available for each child. For this cohort teachers were not given special instructions nor a prescribed curriculum, although most of them used the same curriculum.

3 THE POMDP FRAMEWORK

Almond (2007) provides a generalized model for how a POMDP can represent measurement of a developing proficiency across multiple time points (Figure 1).

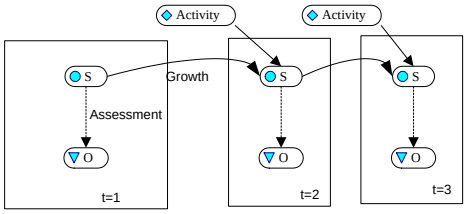

Figure 1: Measurement across time as POMDP

In this figure, the nodes marked **S** represent the latent student proficiency as it evolves over time. At each time slice, there is generally some kind of measurement of student progress represented by the observable outcomes **O**. Note that these may be different for different time slices (c.f., Table 1). Following the terminology of evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003) we call this an *evidence model*. In general, both the proficiency variables at Measurement Occasion m, $\mathbf{S}_m$, and the observable outcome variables on that occasion, $\mathbf{O}_m$ are multivariate.

Extending the ECD terminology, Almond (2007) calls the model for the $\mathbf{S}_m$'s, the *proficiency growth model*. Following the normal logic of POMDPs this is expressed with two parts: the first is the initial proficiency model, which gives the population distribution for proficiency at the first measurement occasion. The second is an action model, which gives a probability distribution for change in proficiency over time that depends on the instructional activity chosen between measurement occasions.

3.1 PROFICIENCY GROWTH MODEL

For the data from the Al Otaiba et al. (2011) study, the latent proficiency is obviously Reading. The question immediately arises as to how many dimensions to use to represent the reading construct. As the students are entering the study in Kindergarten, components of the reading skill, such as oral vocabulary and phonemic awareness are less tightly correlated than they are with older children. (In the fall of the Kindergarten year the correlation between the LW and PV scores in the Al Otaiba et al. study was $r = .46$, $n = 247$, while in the spring it had increased to $r = .56$, $n = 224$.) As

a starting point, we will fit a unidimensional model of Reading, representing it with a single continuous variable: R_{nm} the reading ability of Individual n on Measurement Occasion m.

3.1.1 Model for Growth

In the first cohort of the Al Otaiba et al. (2011) study, teachers were not given specific instructions about curriculum or activity between the time points. We therefore do not have a dependency on activity to measure here. However we do expect there to be some classroom-to-classroom differences, so we will make the growth parameters dependent on the classroom (The teacher effect is part of the classroom effect, however aspects of the peer group and environment are captured as well). Let $c(n)$ be the classroom to which Student n belongs. Note also that classrooms are nested within schools, so school effects are considered part of the general classroom effect.

Following this logic, for Measurement Occasion $m > 1$, define:

$$\begin{aligned} R_{nm} &= R_{n(m-1)} + (\gamma_{c(n)m} + \gamma_{0m})\Delta T_{nm} + \eta_{nm} \\ \eta_{nm} &\sim \mathcal{N}(0, \sigma_{c(n)m}\sqrt{\Delta T_{nm}}) \end{aligned} \tag{1}$$

Here ΔT_{nm} is the time between Measurement Occasions m and $m-1$ for individual n. Here γ_{0m} is an average growth rate, and γ_{cm} is a classroom specific growth rate. Note that the residual standard deviation depends on both a classroom specific rate, σ_{cm}, and the time elapsed between measurements. This is consistent with the model that student ability is growing according to a nonstationairy Wiener (Brownian motion) process.

3.1.2 Model For Initial Proficiency

Children entering Kindergarten have very diverse language and early literacy backgrounds. There are considerable differences in the amount of experience with print material the child experiences at home, breadth and depth of vocabulary used with the child, as well as a wide variety of preschool experiences. As a child's preschool and early home experiences are at least partially dependent on their parents' social and economic status, and within-school socio-economic status tends to be more homogeneous than across school status, we model the initial status as dependent on the school. Let $s(n)$ be the school attended (during Kindergarten) for Student n.

There is also a considerable variation in the age at entry. In the Al Otaiba et al. (2011) study, 95% of the children were between the ages of 5 years 2 months and 6 years 4 months at the time of the first testing (with a few students 7 years or older). This represents a considerable variation in maturity, and potentially in initial ability.

We define the following model for Measurement Occasion 1:

$$R_{n1} \sim \mathcal{N}(\mu_{s(n)}, \upsilon_{s(n)}) \tag{2}$$

3.2 EVIDENCE MODELS

Because we are assuming that Reading proficiency is unidimensional, we do not need to specify which of the measures in Table 1 are relevant to which proficiencies. Thus, the evidence model is a collection of simple regressions, for each observation Y_{nmi} for Individual n at Measurement Occasion m on Instrument i, we have:

$$Y_{nmi} = a_i + b_i R_{nm} + \epsilon_{nmi} \tag{3}$$

$$\epsilon_{nmi} \sim \mathcal{N}(0, \omega_i) \tag{4}$$

Note that the slope parameter b_i actually encodes a relative importance for the various measures.

One advantage of this structure is that we we do not need to explicitly specify the data collection structure (Table 1). Instead, we can simply set the values of measures not recorded in each wave to missing values.

Because each of the instruments are well established (Woodcock et al., 2001; Good & Kaminkski, 2002), we know some of their critical psychometric properties. In particular, the *reliability* of Instrument i, ρ_i is documented in the handbooks for the measures. In classical test theory, the reliability is the squared correlation between the true score of an examinee and the observed score. With a bit of algebra, this definition is is equivalent to:

$$\rho_i = 1 - \frac{\mathrm{Var}_n(\epsilon_{nmi})}{\mathrm{Var}_n(Y_{nmi})} \ . \tag{5}$$

Here the notation $\mathrm{Var}_n(\cdot)$ indicates that the variance is taken over individuals (with measurement occasion and instrument held constant). Solving Equation 5 for $\mathrm{Var}_n(\epsilon_{nmi})$ yields an estimate for ω_i^2 for each measurement occasion. We took the median of the three estimates as our base estimate for ω_i^2, $\tilde{\omega}_i$.

One drawback of the classical test theory concept of reliability is that it is dependent on the population being measured. Thus, as the sample in the Al Otaiba et al. (2011) is slightly different from the norming samples used in the development of the WJ-III and DIBELS measures, we expect our observed reliability will differ slightly from the published values. What we do is set up priors for ω_i using $\tilde{\omega}_i$ as the prior mean. In particular,

$$1/\omega_i^2 \sim \mathrm{Gamma}(\alpha, \alpha\tilde{\omega}_i^2) \ , \tag{6}$$

where Gamma(α, β) is a gamma distribution with shape parameter α and rate parameter β. We note that any gamma distribution with $\beta = \alpha\tilde{\omega}_i^2$ will have the proper mean. The shape parameter α is then effectively a tuning parameter giving the strength the prior distribution, or equivalently the relative weight of the published reliabilities and the observed error distribution. We initially chose a value of $\alpha = 100$ weights the prior knowledge as equivalent to 100 observations, but later increased it to 1000 when we were experiencing convergence problems.

3.3 SCALE IDENTIFICATION

A problem that frequently arises in educational models using latent variables is the identifiability of the scale. In particular, suppose we replaced R_{nmi} with $R'_{nmi} = R_{nmi} + c$ for an arbitrary constant c, and replaced a_i with $a'_i = a_i - b_i c$. The likelihood of the observed data Y_{nmi} (implicit in Equation 3) would be identical. A similar problem arises if we replace R_{nmi} with $R''_{nmi} = cR_{nmi}$ and b_i with $b''_i = b_i/c$. Additional constraints must be added to the model to identify the scale and location of the latent variable R.

A frequently used convention in psychometrics is to identify the scale and location of the latent variable by assuming that the population mean and variance for the latent variable is 0 and 1 (i.e., that the latent variable has an approximately unit normal distribution). In this case we can identify the scale for R_{n1} by constraining $\sum_s \mu_s = 0$ and $\frac{1}{S}\sum_s \upsilon_s = 1$, where S is the total number of schools in the study.

Because this is a temporal model, there exists another complication. We need to identify the scale of R_{nm} for $m > 1$. In particular, the mean and variance of the innovations γ_{0m} and σ_{tm} can cause similar identifiability to the scale and location for R_{nm} that the initial mean and variance caused for R_{nm}. In this case we apply a different solution. We assume that the properties of the instruments, and their relationships to the latent reading proficiency do not vary across time (at least for the time points they are in use). Note that in Equation 3, the slope, b_i and intercept, a_i do not vary across time. This establishes a common scale for all time points.

Our initial thinking was that this would be enough to identify the model. Unfortunately, because of the structural missing data additional constraints are needed. These are described below.

Bafumi, Gelman, Park, and Kaplan (2005) present a different approach to enforcing identifiability. They let the model be unidentified while fitting the data, but then transform the estimates when evaluating the data (i.e., they enforce the constraint by manipulating the samples in R and coda R Development Core Team, 2007; Plummer, Best, Cowles, & Vines, 2006 rather than in BUGS or JAGS). For example, rather than constraining $\sum \mu_s = 0$, they would estimate μ_s freely, but post hoc would adjust the sample from the rth cycle, $\mu_s^{(r)'} = \mu_s^{(r)} - \sum \mu_s^{(r)}$, making appropriate adjustments to the other parameters. They claim that the resulting model mixes better, however, there is some difficulty in figuring out how the post hoc adjustments will affect other parameters in the model.

4 PROBLEMS WITH MODEL FITTING

We attempted to fit the model described in the previous section with MCMC using JAGS (Plummer, 2012).[1] After some initial difficulties we removed the teacher and school effects (intending to add them again after we fit the simpler model). This also allowed us to restrict the prior distribution for R_{n1} to be a unit normal distribution (zero mean, variance one). This is a common identifiability constraint imposed in psychometric models.

4.1 FIVE MEASURE MODEL

Our initial experiments involved five of the six measures (the PSF measure was left out due to a mistake in the model setup). We ran three Markov chains using random starting positions and found that the models did not converge. Or more properly, the evidence model parameters (a_i, b_i, and ω_i) for the DIBELS NWF (nonsense word fluency) measure did not converge. Table 2 shows the posterior mean of the evidence model parameters for the five measures (because the MCMC chain did not reach the stationary state, this may not be the true posterior).

Note in Table 2 that the estimated residual variance is extremely low, indicating a nearly perfect correlation between the latent Reading variable and the NWF measure. In this case, the MCMC chain looks like it is somehow using that measure to identify the scale of the latent variable. Furthermore, the slope for that variables is twice as high as the slope for other variables in

[1] Actually, we did some of our early model fitting using WinBUGS (D. J. Lunn, Thomas, Best, & Spiegelhalter, 2000). Some of the identification problems we were having in WinBUGS we are not having in JAGS. JAGS may be using slightly better samplers which may take care of issues that occur when the predictor variables in regressions are not centered (Plummer, 2012). Similar improvements may have been made in OpenBUGS (D. Lunn, Spiegelhalter, Thomas, & Best, 2009), the successor to WinBUGS, but we have not tested this model using OpenBUGS.

Table 2: Evidence Model Parameters, 5 Measure Model

	LW	PV	LNF	ISF	NWF
a	105.37	99.90	25.76	13.97	-4.27*
b	0.15	0.05	0.49	0.32	0.87*
ω	6.15	4.92	6.31	4.38	0.09

* indicates parameter did not converge

the model. Table 3 shows some of the difficulty. The NWF measure is the only one showing a large increase between the Winter and Spring testing periods. So naturally, there is a tendency to track that measure.

Table 3: Mean Scores on Each Measure at Each Administration

	LW	PV	LNF	ISF	NWF
Fall	108.5	100.6	27.3	14.2	
Winter	110.8	102.3	42.9	25.5	27.9
Spring	111.2	101.7	51.3		43.2

Trace plots of the evidence models show the problem. Figure 2 shows an example of extremely slow mixing, that is characteristic of identifiability problems. Depending on the values of the other variables in the system (particularly the latent reading variables) higher or lower slopes may be sensible. Looking at the trace plots of R_{nm} for several students show similar poor mixing for $m > 1$. We would expect similar problems with the trace plots for γ_{0m}, but the mixing looks good on those chains.

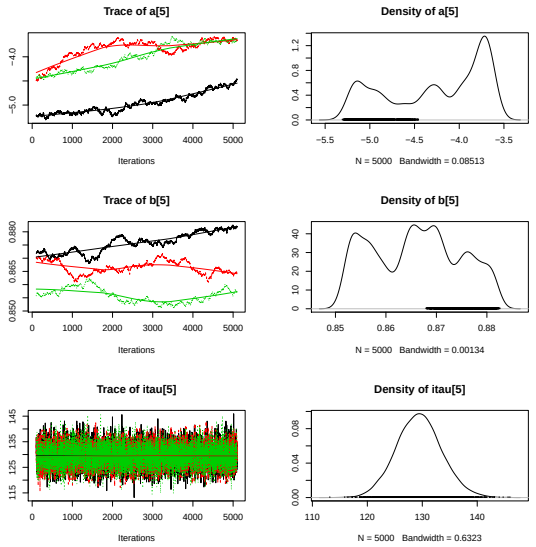

Figure 2: Trace plots of evidence model parameters for measure NWF

It is likely that the problem is some complex interaction between using γ_{0m} and the b's to identify mean growth, or the a's which define the starting point for growth. Note that the problematic measure, NWF, was not measured at the first time point. Thus the constraint on the distribution of R_{n0} will not define its scale in the second or third measurement occasions.

4.2 THREE MEASURE MODEL

As the problematic measure may be the ones which were not recored at all three time points, we ran the model again, dropping the ISF and NWF measures (the ones not observed at the first or third measurement occasion). The new model also did not converge, although the focus of the problem has now moved from the NWF measure to the LNF measure.

Table 4 shows the new estimates from the unconverged posterior. Again, the variance for the measure that did not converge is substantially smaller than that of the other measures, and the slope is substantially higher. Again the trace plots (Figure 3) show poor mixing, as do similar plots for the R_{nm} measures for $m > 1$. There is also an indication of a trend that indicates that the chains have not covered the whole of the posterior distribution.

Table 4: Evidence Model Parameters, 3 Measure Model

	LW	PV	LNF
a	103.74	98.95	23.02*
b	1.59	0.64	4.55*
ω	5.55	4.78	0.11

* indicates parameter did not converge

4.3 MISSING IDENTIFICATION CONSTRAINT

Looking back to the problems in the model fit in Section 4.1, note that the lack of fit could be explained by the interaction between a_5 (the intercept for the NWF measure) and γ_1 (the average proficiency change between the first and second time points). As NWF is not measured in the first time point, any arbitrary change between the first and second time point can be created by changing a_5, b_5 and γ_1. The other four measures were all collected in the fall, so in these cases, a_i should have been fixed by the constraint that $E[R_{n1}] = 0$.

What is required is a method for fixing the value of a_i for measures that were not collected at the initial time point. One possible way to do this would be to simply set $a_i = 0$. This is not unreasonable, if all of

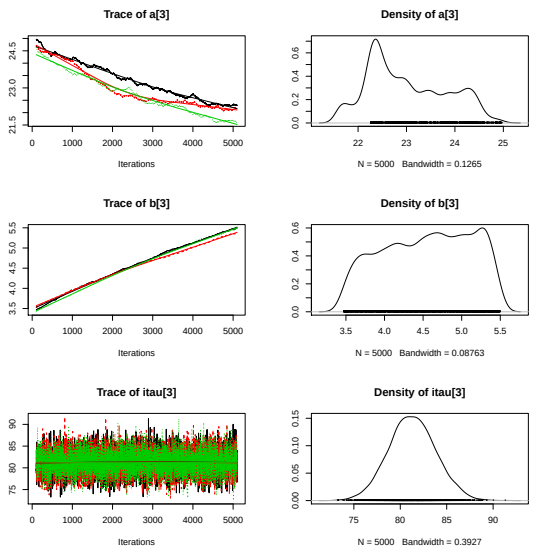

Figure 3: Trace plots of evidence model parameters for measure LNF, Three Measure Model

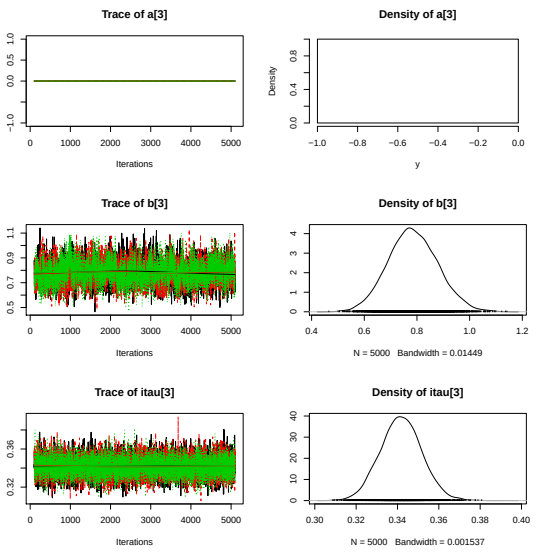

Figure 4: Trace plots of evidence model parameters for measure NWF, Six Measure Model with $a_i = 0$

the variables are on a standardized scale: it implies that the average trajectory of the average student will pass through the average of the scores.

This required that the scores all be on the same scale (especially problematic with the WJ-III and DIBELS scores based on different development and norming sample. Fortunately, for these data all six measures were collected in the winter time period. Subtracting the mean of the Winter scores and dividing by the standard deviation for each measure produced standardized scores. This standardization together with the constraint $a_i = 0$ caused the models to converge.

4.4 SIX MEASURE MODEL

Using the standardized data and the additional constraint of $a_1 = 0$, we again fit the model using MCMC. This time, we got convergence on all of the evidence model parameters (Figure 4).

Table 5 shows the mean of the latent Reading variable for the first five students in the sample. This appears to be well behaved with all of the students showing growth across the three time points.

Table 5: Mean values for Reading for first five students, Six Measure model with $a_i = 0$

	S1	S2	S3	S4	S5
F	-0.394	-0.351	-0.375	-1.556	-0.773
W	0.029	0.104	0.031	-1.278	-0.415
S	0.864	1.016	0.852	-0.514	0.419

5 FUTURE DIRECTIONS AND CHALLENGES

The key to getting this model to converge was the standardization of the measure scales. Fortunately, this data set had a time period where all six measures were applied to the same population. Consequently, standardizing the scales at this time point put the measures on a comparable scale, which then made the fixed intercept constraint meaningful.

It is difficult to see how this generalizes to cases in which there is not a single time point in which all measures are collected. This is a problem with the cohort examined in this study when we look at the data gathered in first and second grades. As the students reading abilities develop, new and more difficult measures of reading become appropriate. Linking these back to the old scale is a difficult problem. This problem is well known in the educational literature under the name "vertical scaling" (von Davier, Carstensen and von Davier, 2006, provide a review of the literature).

Now that the model without teacher or school effects converges, the next step is to add those back into the model. Also, we should use cross-validation to evaluate how well the model predicts students scores. The Al Otaiba et al. (2011) data set has long term follow-up for a substantial portion of the students, so we can see how well the model can predict First and Second grade reading scores as well. Finally, we can look at the rules for classification in to special instruction, to see whether integrating the data across multiple measures

provides a better picture of the student than looking at one measure alone.

Acknowledgments

We would like to thank the Florida Center for Reading Research for allowing us access to the data used in this paper. The data were originally collected as part of a larger National Institute of Child Health and Human Development Early Child Care Research Network study.

References

Almond, R. G. (2007). Cognitive modeling to represent growth (learning) using Markov decision processes. *Technology, Instruction, Cognition and Learning (TICL)*, *5*, 313–324. Available from `http://www.oldcitypublishing.com/TICL/TICL.html`

Almond, R. G. (2009). *Estimating parameters of periodic assessment models* (Research Report No. To appear). Educational Testing Service.

Almond, R. G. (2010). Using evidence centered design to think about assessments. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs.* (pp. 75–100). Springer.

Al Otaiba, S., Folsom, J. S., Schatschnneider, C., Wanzek, J., Greulich, L., Meadows, J., et al. (2011). Predicting first-grade reading performance from kindergarten response to tier 1 instruction. *Exceptional Children*, *77*(4), 453–470.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practial issues in implementing and understanding bayesian ideal point estimation. *Political Analysis*, *13*, 171-187.

Bennett, R. E. (2007, May). *Assessment of, for, and as learning: Can we have all three?* Paper presented at the Institute of Educational Assessors National Conference, London, England. Available from `http://www.ioea.org.uk/Home/news_and_events/annual_conference/day1/randy_bennett.aspx`

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*(1), 7–74.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, *11*, 1-94. Available from `citeseer.ist.psu.edu/boutilier99decisiontheoretic.html`

Good, R. H., & Kaminkski, R. A. (Eds.). (2002). Dynamic indicators of basic early literacy skills (6th ed.) [Computer software manual]. Available from `https://dibels.uoregon.edu/`

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049–3082.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*, 315-335.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3-62.

Pelligrino, J., Glaser, R., & Chudowsky, N. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* National Research Council.

Plummer, M. (2012, May). JAGS version 3.2.0 user manual (3.2.0 ed.) [Computer software manual]. Available from `http://mcmc-jags.sourceforge.net/`

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). coda: Output analysis and diagnostics for MCMC [Computer software manual]. (R package version 0.10-7)

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Raudenbush, S. W., & Byrk, A. S. (2002). *Hierarchical linear models* (second edition ed.). Sage Publications.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence* (1st ed.). Oxford University Press, USA.

von Davier, M., Carstensen, C. H., & von Davier, A. A. (2006). *Linking competencies in educational settings and measuring growth* (Research Report No. RR-06-12). ETS.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* Jossey-Bass.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Wj-iii tests of cognitive abilities and achievement [Computer software manual].