Scoring and then analyzing or analyzing while scoring: An application of GLMM to an education instrument development and analysis

Mark C. Greenwood * Dan Jesse†

Abstract

Psychometric methods provide clear ideas for mapping scored (binary) items into overall test scores using Rasch or Item Response Theory models or, more simply, scores can be based on percentages. We explore potential analysis techniques for longitudinal responses when large, randomly sampled test data sets may not be available, merging the scoring and analysis processes into a single model. Specifically, logistic generalized linear mixed models are used to analyze the binary item responses for scoring information (item difficulty and subject scores) and to address longitudinal research questions in the same model. This approach is compared to a generalized linear mixed model analysis of the count of items correct and a linear mixed model analysis of the one-parameter logistic model score results based on an independent data set. The different methods are motivated by the development and analysis of an instrument from a five-year longitudinal study of coaches of K-8 mathematics teachers. In this application, there are negligible differences in using the three different approaches.

Key Words: IRT, GLMM, Longitudinal model, Psychometric, Education data

1. Introduction

It is extremely common for statistical results in educational research to be based on the analysis of scores developed through Psychometric methods. These scores can be developed with varying methods with the Item Response Theory (IRT) methods providing a model-based framework to move from the original, binary, correct or not, responses on all the individual items in instrument (or test) to scores that are estimates of the underlying latent ability of subjects (de Ayala, 2008, among others). In a longitudinal study, either the first time of observation or, more typically, a separate data set must be used to perform the IRT analysis to develop a scoring model and then the longitudinal responses are scored using that model. In these studies, the number of subjects may not be large enough to develop the scoring model from the first time point and it may be suspect to use all of the responses for developing the scoring model since their change over time might be of interest in the research. If researchers cannot appeal to prior work for scoring an instrument, as in the study of interest where the instrument was developed as part of the longitudinal study, an independent data set may not be available or it may not be "large". In this study, a moderate sized, independent data set was available (n=191) and the results from developing a scoring model using it are compared to focusing exclusively on the longitudinal responses. The connections between IRT and generalized linear mixed models (GLMM) are exploited to connect and compare the different approaches. Specifically, modeling the binary item-level responses for each subject over time is compared to analyzing the binomial counts over time, and both are compared to analyzing the IRT-scores using a linear mixed model. For the reasons discussed below, there is essentially no difference in the conclusions across the three different analysis approaches.

The paper provides an outline of the study and the source of the data sets, the scoring methods employed and the associated linear mixed model analysis of those results, a dis-

*Montana State University, Bozeman, MT, 59717

†RMC Research Corporation, Denver, CO

cussion of scoring and analysis using a binomial generalized linear mixed model that is not typically used in these situations, and an item-level analysis of the binary responses. In conclusion, some general recommendations based on this study are provided.

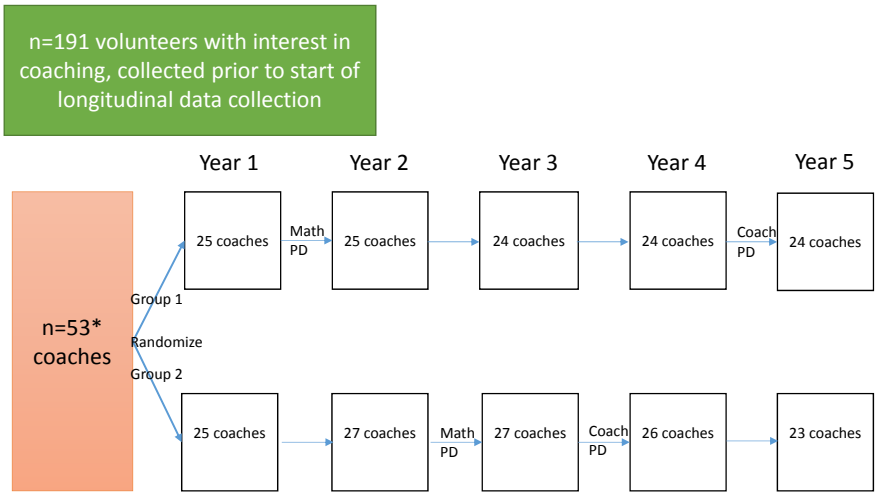
1.1 Study and data collection

A coach works collaboratively with a teacher to improve the teacher's practice and content (here mathematics) knowledge (Sutton, Burroughs, Yopp, 2011). While vast resources have been expended on coaching programs in the United States, there is limited research available on the effectiveness of coaching. Similarly, there is little information on what characteristics describe effective coaches. One potential aspect of coaching is knowledge of what is recommended in the literature related to coaching methods. The Examining Mathematics Coaching project was funded by the National Science Foundation (NSF Discovery Research K-12 Program, Award No. 0918326; `http://www.math.montana.edu/~emc`) to study the impacts of the knowledge of elementary/middle school coaches on the way mathematics teachers teach. In order to explore the potential linkage between coaching knowledge and teacher practice, we needed to develop a measure of coaching. Specifically, we needed to develop a measure of coaching knowledge to provide for measurements of impacts of the randomly assigned professional development (see Figure 1 for the timing of the training experiences in the two-armed study) and for assessing whether variation in coaching knowledge was related to variation in measures of the coaches' teachers (not explored here). There were no existing instruments to measure this knowledge and so the researchers used a modified Delphi method with a panel of twelve experts to develop a large set of items related to the coaching literature. More details of the instrument development are available in Sutton et al. (2011). As part of this process, a group of 191 volunteers agreed to take the instrument, solicited from those who participated in a related interest group and had general expertise in research related to coaching. Responses from these subjects provide a data set, called the "pilot" data set below, to explore the items and estimate the initial scoring model. This scoring model can then be employed to the longitudinal responses of 53 different participants in the longitudinal mathematics coaching project who completed measures as described in Figure 1.

As described in Sutton et al. (2011), a modified Delphi method was used with twelve panelists to construct items that measure various aspects of recommendations for coaching in the coaching literature. The items developed were mostly Likert scales from 1 (disagree) to 7 (agree) although a few multiple choice items were constructed that contained four potential responses. In order to develop an overall score for "coaching knowledge", researchers established "correct" responses based on selections of responses on the Likert scales or the multiple choice items. In some cases, it was not possible to agree on whether there was a correct response and those items were dropped from further analyses. This process retained 40 items that were available for the next steps in the analysis.

2. Scoring and then analyzing scores

One of the main advantages of having an independent data set for calibrating the scoring model is that it provides a framework for assessing the dimensionality of the items in an instrument and for refining the items used, especially if there are items that do not agree with the overall pattern of other responses. Additionally, it is possible to obtain information on the difficulty and potential differences in item discrimination. More information on the IRT process can be found in de Ayala (2008) and about using R (R Core Team, 2014) for these purposes in Rizopoulos (2006). Generally, a single underlying latent trait is sought



* Coaches were randomly assigned to groups and replaced with new coaches prior to year 2.

Figure 1. Schematic of study design and sample sizes available at different times.

that measures the latent ability or knowledge trait in the subjects (Reckase, 2009) that is based on a large number of items. If items do not conform to this single trait, then they can be considered for removal from the model. A combination of exploratory factor analysis, hypothesis testing for unidimensionality, and fitting of different types of IRT models were used here to refine the items used until a single set of items and scoring model are converged on. A description of the application of this process to the pilot data set follows.

Initially, 40 items were identified from the original instrument from which the researchers could agree on the "correct" answer - the answer that conformed with the recommendations in the coaching literature. Of those items, one had 100% "correct" responses in the pilot data set. While it might not have had the same results in the longitudinal data set, that item was removed because of potential issues with separation it could create in the IRT models considered. The remaining 39 items were then used in the following analyses. In order to assess the dimensionality of those responses, a one-factor, maximum likelihood exploratory factor analysis of the tetrachoric correlation matrix of was performed (Revelle, 2014). Because the tetrachoric correlation matrix was not full rank, it had to be smoothed to be invertible (Revelle, 2014). In this analysis, there were 20 items with loadings over 0.4 on the single factor (Table 1). Additional attempts to explore higher dimensional solutions resulted in factors with few traits and some negative loadings - suggesting that the latent trait was something other than coaching knowledge, that some aspects of coaching knowledge may be contradictory or based on factions within the literature, or that the items were flawed in some way. In order to retain items that provided a single "pure" knowledge trait, only the items that loaded on the initial latent trait were considered in the IRT models that follow. In all remaining administrations of the instrument, the subjects took all items but only the responses on these questions were utilized.

Table 1: Results from preliminary EFA of tetrachoric correlation of initial 39 items.

Item	Standardized Loading	Communality	Uniqueness	Retained
item19	0.80	0.64	0.36	Yes
item16	0.74	0.55	0.45	Yes
item26	0.64	0.41	0.59	Yes
item30	0.64	0.41	0.59	Yes
item18	0.62	0.38	0.62	Yes
item27	0.61	0.37	0.63	Yes
item31	0.61	0.37	0.63	Yes
item32	0.61	0.37	0.63	Yes
item11	0.58	0.34	0.66	Yes
item17	0.58	0.34	0.66	Yes
item33	0.57	0.33	0.67	Yes
item21	0.55	0.30	0.70	Yes
item34	0.54	0.29	0.71	Yes
item22	0.50	0.25	0.75	Yes
item28	0.49	0.24	0.76	Yes
item14	0.48	0.23	0.77	Yes
item15	0.48	0.23	0.77	Yes
item10	0.46	0.21	0.79	Yes
item12	0.44	0.20	0.80	Yes
item20	0.43	0.19	0.81	Yes
item23	0.33	0.11	0.89	No
item7	0.27	0.08	0.92	No
item24	0.27	0.07	0.93	No
item4	0.25	0.07	0.94	No
item5	0.17	0.03	0.97	No
item2	0.16	0.03	0.97	No
item36	0.15	0.02	0.98	No
item29	0.12	0.02	0.98	No
item35	0.11	0.01	0.99	No
item1	0.07	0.00	1.00	No
item8	0.05	0.00	1.00	No
item3	0.04	0.00	1.00	No
item13	0.04	0.00	1.00	No
item39	0.04	0.00	1.00	No
item38	0.00	0.00	1.00	No
item9	-0.06	0.00	1.00	No
item37	-0.09	0.01	0.99	No
item25	-0.14	0.02	0.98	No
item6	-0.17	0.03	0.97	No

2.1 IRT methods and results

With the 20 items identified as likely to be related to a single dimensional latent trait, the other 18 items were discarded. If one allows the use of the full suite of IRT models, there is the potential to estimate scoring models that range from containing just different item difficulties and a fixed, common discrimination of items to models that contain different difficulties, discrimination, and guessing parameters for each item (de Ayala, 2008). These more complex models require extremely large sample sizes to be confidently estimated and so the focus here will be on the simpler IRT models. Because the subjects in the pilot data set were generally knowledgeable about the coaching literature, guessing was not a likely component of their responses. This may not have been the case for the coaches in the longitudinal data set but the sample size precluded trying to incorporate that component in the models.

The IRT models map patterns of multivariate binary responses into continuous latent "ability" traits, z_m, for all m subjects (Rizopoulos, 2006). The so-called 2-parameter logistic model for the mean of m^{th} subject on the i^{th} item is $logit(P(x_{im} = 1|z_m)) = \beta_{1i}(z_m - \beta^*_{0i})$ where z_m is the underlying latent "ability" trait for subject m, β_{1i} is the discrimination parameter (slope and the focus of the item on its particular difficulty level), and β^*_{0i} is the difficulty parameter (intercept and center of discrimination region for this question). Higher discrimination values relate to more focused information from that item and higher difficulty level items discriminate subjects with higher latent abilities. The three standard IRT models are the Rasch, the 1-parameter logistic, and the 2-parameter logistic models. The Rasch model fixes β_{1i} at 1 for all items, but allows different item difficulty levels. The 1-parameter logistic(1 PL) model involves a common discrimination parameter $\beta_{1i} = \beta_1$ for all i, but it is estimated instead of being fixed at 1. The 2-parameter logistic model has different difficulty and discrimination as the initial parameterization suggested.

Because each model is nested within the next more complicated model, hypothesis testing can provide information about the need for the additional complexity; with a modest sample size, it may be difficult to find evidence for the most complicated of the models considered. Specifically, likelihood ratio tests are used to assess evidence for adequacy of each simpler model, starting with the Rasch model up to the 2 PL. For a selected model, it is also important to check for overall lack of fit and unidimensionality, both possible using bootstrap tests (Rizopoulos, 2006).

In the 1 PL model, the discrimination parameter is estimated to be 1.24 with an SE of 0.096 and the likelihood ratio test provides strong evidence against the null hypothesis that the Rasch model is adequate (χ^2_1=7.86, p-value=0.005). There is only moderate evidence that the 2 PL is needed when compared to the 1 PL (χ^2_{19}=29.4, p-value=0.06). Because of this ambivalent evidence in support of the much more complicated model, the 1 PL is retained. There is no evidence of a lack of fit of the 1 PL model from a bootstrap test with a p-value of 0.556, where the null hypothesis was of model adequacy with the 1 PL model. Similarly, the test for unidimensionality, where null hypothesis is that the second eigenvalue of the data is as expected under the 1 PL model and the alternative is that it is greater than expected, provided a p-value of 0.204 from 999 bootstrap samples. This last result is not too much of a surprise since the items passed to the 1 PL model were selected based on being highly associated with the latent trait in the exploratory factor analysis, but this result is nonetheless reassuring.

Item characteristic curves (ICCs) plot the probability of a subject getting a response correct on an item as a function of the estimated latent ability and the item information curves (IICs) display the derivative of the ICCs and more clearly show where each item is able to discriminate among subjects with different latent abilities. Comparing the ICCs and

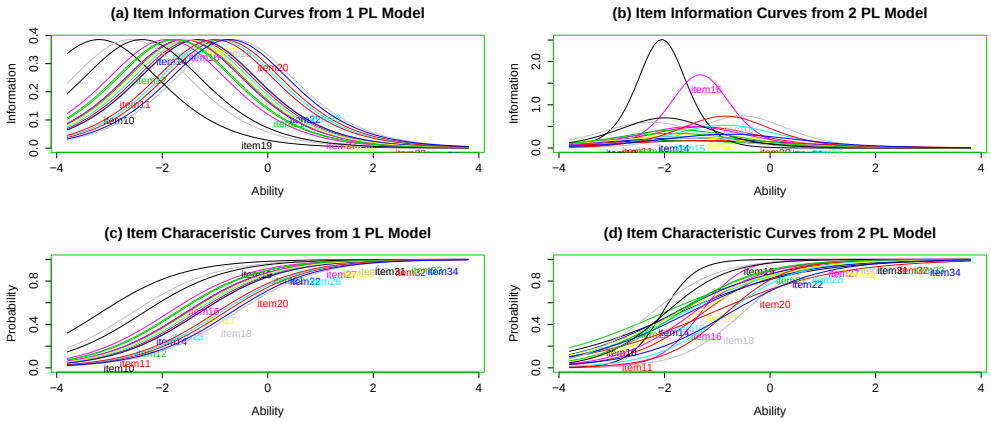


Figure 2: Plots of estimated IICs and ICCs for the 1 PL and 2 PL models.

IICs for the 1 and 2 PL models aids in understanding the previous comparisons of the need for additional model complexity. This comparison also provides information for assessing potential lack of the fit of the 1 PL model by allowing the discrimination parameters to be estimated for each item. Exploration of the curves for the 1 PL and 2 PL models in Figure 2 shows that there is little difference in discrimination among the different items in the 2 PL model. It also shows that most of the items were relatively easy for the subjects in the pilot data set and focused their discrimination on the lower ability subjects. Harder items would have IICs centered at higher abilities and would have more clearly discriminated the differences in the higher latent ability subjects. It ends up that the subjects in the longitudinal study started at the lower ability levels and so the discrimination in this lower region may have been useful in this particular study.

2.2 Longitudinal Analysis

With the 1 PL selected and estimated based on the pilot data set, it is possible to estimate scores for the subjects in the longitudinal study, which we will call Coaching Knowledge Survey (CKS) scores. The resulting estimated scores for these subjects (Figure 3) shows that they started, on average, about 1 standard deviation of the CKS latent ability trait below the group used in the pilot data set. Over the course of the study, the coaches in the study approached the mean of 0 generated from the pilot data set. The difference between PD Groups 1 and 2 is minor visually but a model that accounts for the repeated measures nature of the measurements is required to fully assess those potential differences. Additionally, we need to control for other potential sources of training that could have impacted the coaching knowledge of the subjects. We measured the presence of outside training prior to the project (Year 1) and in the twelve months prior (Years 2 through 5) in coaching and in mathematics, creating binary predictor variables of 0: no outside training and 1: outside training. A linear mixed effects model (Pinheiro et al., 2013) for i^{th} CKS score on j^{th} coach (CKS$_{ij}$) is (loosely) written as $CKS_{ij} = Time_{ij} * PDGroup_i + OutsideCoachPD_{ij} + OutsideMathPD_{ij} + \mathbf{Coach_j} + \epsilon_{\mathbf{ij}}$. The main features of this model are that it contains a random coach effect, $\mathbf{Coach_j} \sim N(0, \sigma^2_{Coach})$, time by PD fixed effects, and control variables for outside coaching and mathematics PD prior to the current year's observation.

There was a relatively high intra-class correlation in the CKS scores, with a model-based estimate of 0.61 for the correlation in two measurements of the same coach once the fixed effects were accounted for. There was no evidence of a time by PD interaction (F(4,189)=0.66, p-value=0.62). If the interaction is removed from the model, there was

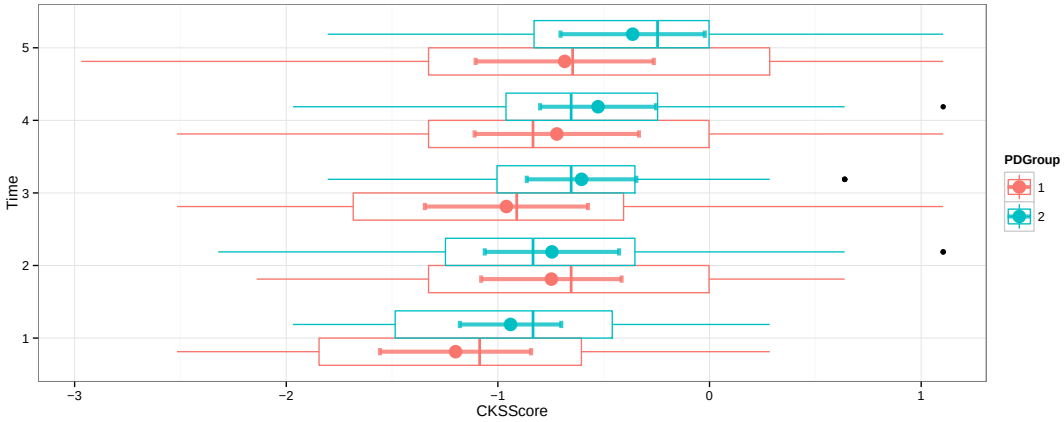


Figure 3: Boxplots with means and approximate 95% confidence intervals of CKS scores over time by PD Group.

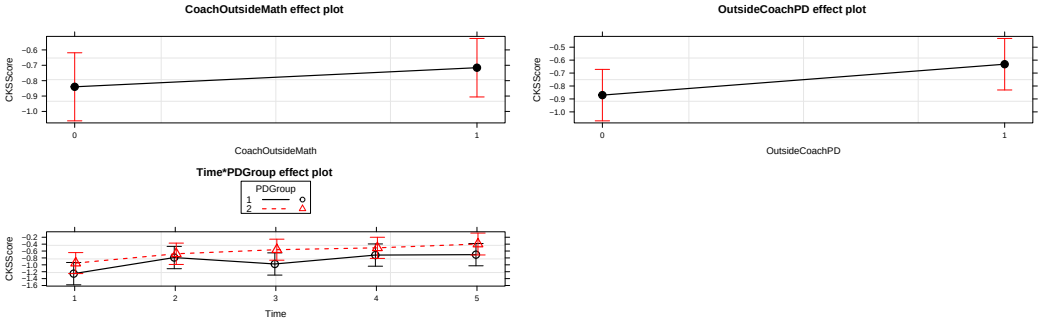


Figure 4: Effect plots from linear mixed model of CKS Scores with model-based 95% confidence intervals.

strong evidence for changes over time ($F(4,193)=8.8$, p-value<0.0001) and impacts of outside coaching PD ($F(1,193)=10.1$, p-value$=0.002$) but no evidence of a difference in the two PD groups ($F(1,49)=2.1$, p-value$=0.153$) or based on having had outside mathematics training ($F(1,193)=1.96$, p-value$=0.163$). Both the time and outside coaching effects were estimated to be in the direction of positive changes on the mean CKS scores. Figure 4 displays the estimated effects from this model, with the y-axis scaling on the 1 PL estimated latent trait scale from the pilot data set. More details of this model are discussed below.

3. Binomial response scoring and analysis

Another approach to modeling data sets that are regularly subjected to IRT analysis is to just count the number of successes out of the total number of items for each subject. This produces a binomial count with the number of items providing the number of trials in the Binomial. It collapses the responses across the items but still tracks the overall success rate. There is no clear way to include the pilot data set, since it is only measured once, into this model. These results focus internally on the longitudinal subjects results only. As in the previous linear mixed model analysis, there are repeated measures for each subject across time to account for. In this model, there are two ways to produce "scores" for each observation on the coach at each time from a model like this. One could include a nested set of random intercepts to account for the differences among subjects and differences in responses over time. The other option would be to create a single level random intercept

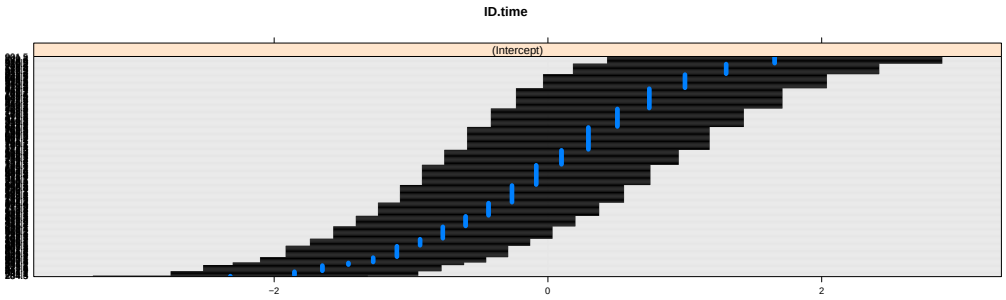


Figure 5: Caterpillar plot of the random effect point estimates from the binomial scoring model with 95% posterior probability intervals.

that provides a unique estimate for each combination of time and subject. The first option would require obtaining estimates from the two levels of random effects and combining them to generate an estimate for the subject at each time. The second option only involves a single random effect and so makes obtaining scores for each subject at each time easier and is used here.

The "scoring" model for the binomial counts is a generalized linear mixed model (GLMM) estimated using glmer (Bates et al, 2014) for the count (Y_{ij}) of conforming items out of 20 is $Y_{ij} \sim Binomial(20, \pi_{ij})$ for i^{th} time on j^{th} coach with $logit(\pi_{ij}) = \beta_0 + \mathbf{Time.Coach_{ij}}$. The random effect, $\mathbf{Time.Coach_{ij}} \sim N(0, \sigma^2_{Time.Coach})$, is estimated by defining a unique level for each administration time for each coach. The random effect can only take on values based on getting 0, 1, 2, ..., 20 correct responses so will be relatively discrete on the log-odds scale (Figure 5). The intercept coefficient, β_0, is the log-odds mean of those responses in this model. This model is very similar to just using the logits of observed proportions correct except that it provides slightly different fitted values because of the assumption that $\mathbf{Time.Coach_{ij}} \sim N(0, \sigma^2_{Time.Coach})$. Note that this model ignores the repeated measures on coaches but also provides simple access to scores for the coaches at each time.

To define an "analysis" model, one needs to account for differences among the subjects with a subject-level random effect and use fixed effects like those in the linear mixed model. Specifically, the model for the logit of the mean response is $logit(\pi_{ij}) = Time_{ij} * PDGroup_i + OutsideCoachPD_{ij} + OutsideMathPD_{ij} + \mathbf{Coach_j}$, with $\mathbf{Coach_j} \sim N(0, \sigma^2_{Coach})$ to account for repeated measures on each coach. Based on this model, there is no evidence of a time by PD interaction (χ^2_4=6.4, p-val=0.17). After removing the interaction effect, there was strong evidence for changes over time (χ^2_4=51.4, p-value<0.0001) and impacts of outside coaching PD (χ^2_1=18.4, p-value<0.0001) but no evidence of a difference in the PD groups (χ^2_1=2.21, p-value=0.137) or of an effect of outside mathematics training (χ^2_1=1.51, p-value=0.219). Similar to the previous linear mixed model, the effects of time and outside coaching PD are also both in the direction of higher probabilities of conforming responses. The estimated effects from this model are provided in Figure 6.

These results are quite similar to those found in Figure 4 except that all interpretations are now in terms of probabilities (or log-odds) of getting a correct response. While this scale has some potential interpretation benefit, it also never references the outside standard provided by the scoring system developed using the pilot sample. It also requires interpretations on a nonlinear scale due to the logit link. Another major difference in these results is the lack of item-level difficulty information. For a researcher, knowing which items are more or less difficult can be quite useful information and this model can never distinguish that level of detail, even though its inferences are similar to those from the linear mixed

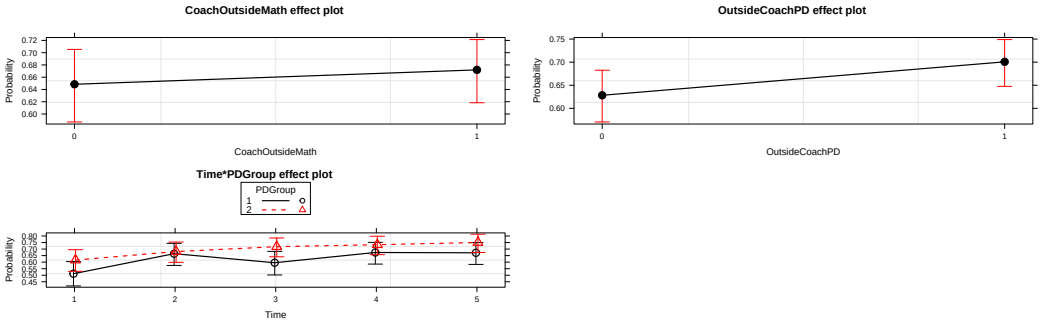


Figure 6: Effect plots from the Binomial-response GLMM analysis with 95% confidence intervals.

model of the CKS scores.

4. Item-level response scoring and analysis

One final approach is possible by focusing on the binary, item-level, longitudinal responses, accounting for multiple measures per time and over time for each subject by employing a multilevel generalized linear mixed model. This idea was inspired by the multi-level Rasch models proposed in De Boeck et al. (2011) and Doran et al. (2007). Both papers emphasize the connections between Rasch models and GLMM and using lme4's glmer (Bates et al., 2014) to both estimate item difficulty and subject latent ability within a single GLMM. Their applications focus on hierarchical data instead of longitudinal data, which is a minor difference except when there are expectations of subjects being different in later times than at the beginning of the study. The main advantage of connecting Rasch models to GLMM is the ability to handle more complex data structures; here the GLMM is used to account for repeated measures at two levels on subjects. Another issue raised in these papers is whether the item effects are treated as fixed or random effects. De Boeck et al. (2011) treat the item levels as fixed effects, assuming that only these items are of interest. Doran et al. (2007) uses a random effect for the items. With a large number of items, it often is reasonable to treat the items as a random sample of potential levels that follow a normal distribution but this assumption can impact the estimated item difficulties. In standard Rasch models, the item difficulties are not typically assumed to follow a particular distribution so it likely is more typical to use fixed effects for the item difficulties. If item levels are random, they would need to be modeled as crossed random effects with the data collection structure - a particular item means the same thing for all subjects and is not unique for a particular subject at a particular time as would be implied by nesting items in observation times. One other benefit of being able to use GLMM for Rasch-modeling situations is one can directly incorporate covariates in the model. In this situation, we used time and training effects as "external" covariates, using the terminology from De Boeck et al. (2011).

The scoring model built from the item-level binary responses includes a nested random effect for the time of each observation, which groups the responses on all the items at a given time relative to their respective subject. Specifically, this model for the i^{th} measurement time on j^{th} subject for the k^{th} item is $Y_{ijk} \sim Binomial(m = 1, \pi_{ijk})$ with $logit(\pi_{ijk}) = \beta_1 Item_1 + ... + \beta_{20} Item_{20} + \mathbf{Time.Coach_{ij}}$. Again, scores for each time for each subject are available from the estimated random effect. Now, the slopes for the item variables provide log-odds scale estimates of the item difficulties. This model was parameterized with fixed effects in a "cell means" fashion (Greenwood and Banner, 2014) so that these results are directly available in the model output. The estimated item effects

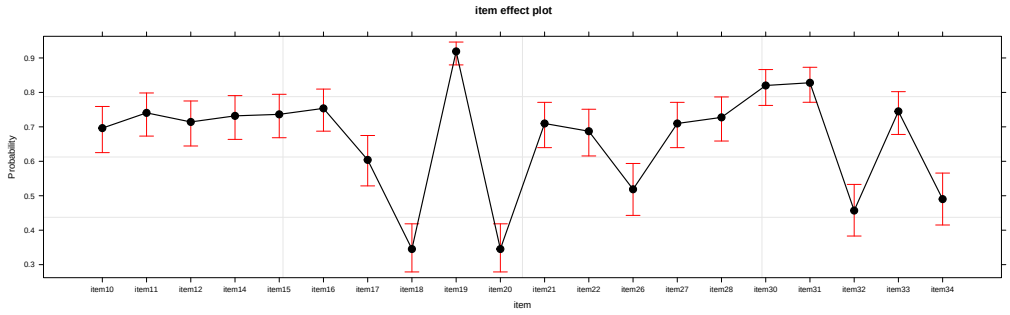


Figure 7: Effect plot for the item difficulties on the probability scale from the Binary scoring model.

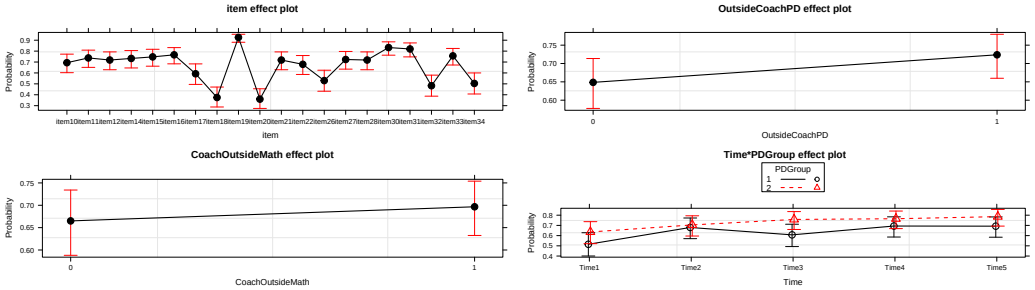


Figure 8: Effect plots from the Binary-response GLMM analysis with 95% confidence intervals.

on the probability scale are displayed in Figure 7. A comparison of the estimated item-specific coefficients from the 1 PL model estimated with the pilot data set and the binary GLMM model from the longitudinal data is provided in Table 2. While there is not perfect agreement in the estimated item difficulties, there are strong similarities in the relative difficulties of the items. This provides an explanation for the similarities observed in results across the different approaches considered here.

The analysis model for the mean is a three-level GLMM with $logit(\pi_{ijk}) = Item_1 + \ldots + Item_{20} + Time_{ij} * PDGroup_i + OutsideCoachPD_{ij} + OutsideMathPD_{ij} + \mathbf{Coach_j} + \mathbf{Time.Coach_{ij}}$, noting that Time.Coach is now a nested random effect relative to the random Coach effect. In this model, the typical R model parameterization with a baseline level and deviations from that level are used even though the model notation used here does not reflect this. In this model, there continues to be no evidence of an interaction between Time and PD ($\chi^2_4 = 4.3$, p-val=0.37). With the interaction removed, there is strong evidence for differences in items ($\chi^2_{19} = 353.4$, p-val<0.0001), outside coach training ($\chi^2_1 = 9.9$, p-val=0.002), and Time ($\chi^2_4 = 30.5$, p-val<0.0001) but little evidence of a difference in the PD groups ($\chi^2_1 = 2.37$, p-val=0.124) or of an effect of outside mathematics training ($\chi^2_1 = 1.31$, p-val=0.252). The strong evidence for an item effect is not a surprise as one would expect some items to be easier than others. In this model, the random effects are not "scores" but included to control for three-level data collection structure.

5. Conclusions

This paper presented three different approaches to obtaining scores and assessing impacts over time on coaching knowledge. All three approaches provided similar results which

Table 2: Item difficulty and log-odds of success from pilot and longitudinal models, with items sorted from easiest to hardest based on estimated difficulties (1 PL) and log-odds of success (Binary GLMM).

Pilot data	1 PL		Long. data	Binary GLMM	
Item	Difficulty	SE	Item	log-odds	SE
Item 19	-3.197	0.398	Item 19	2.429	0.224
Item 30	-2.597	0.311	Item 31	1.572	0.181
Item 31	-2.386	0.286	Item 30	1.517	0.180
Item 16	-1.910	0.238	Item 16	1.118	0.168
Item 33	-1.782	0.227	Item 33	1.072	0.167
Item 21	-1.742	0.224	Item 11	1.049	0.167
Item 14	-1.591	0.212	Item 15	1.027	0.166
Item 27	-1.590	0.212	Item 14	1.004	0.166
Item 12	-1.555	0.209	Item 28	0.982	0.165
Item 28	-1.520	0.207	Item 12	0.916	0.164
Item 15	-1.453	0.202	Item 21	0.894	0.164
Item 11	-1.387	0.197	Item 27	0.894	0.164
Item 22	-1.325	0.193	Item 10	0.830	0.162
Item 10	-1.325	0.193	Item 22	0.788	0.162
Item 17	-1.325	0.193	Item 17	0.423	0.157
Item 32	-1.010	0.176	Item 26	0.075	0.155
Item 26	-0.931	0.172	Item 34	-0.039	0.155
Item 20	-0.854	0.169	Item 32	-0.172	0.155
Item 34	-0.731	0.164	Item 18	-0.640	0.159
Item 18	-0.635	0.161	Item 20	-0.640	0.159

Table 3: Estimated model coefficients from the three analysis models. The estimated intercept from the Binary GLMM (*) also is aliased with the estimate of the log-odds of item 10.

	Linear MM		Binomial GLMM		Binary GLMM	
Effect	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	-1.462	0.185	-0.189	0.211	-0.189*	0.306
Time[2]	0.467	0.146	0.638	0.146	0.692	0.207
Time[3]	0.280	0.146	0.340	0.144	0.376	0.205
Time[4]	0.539	0.147	0.679	0.147	0.757	0.209
Time[5]	0.551	0.146	0.668	0.146	0.751	0.208
PDGroup[2]	0.306	0.226	0.426	0.264	0.496	0.338
CoachOutsideMath	0.124	0.089	0.105	0.090	0.146	0.128
OutsideCoachPD	0.239	0.079	0.326	0.079	0.350	0.115
Time[2]:PDGroup[2]	-0.197	0.201	-0.352	0.197	-0.379	0.292
Time[3]:PDGroup[2]	0.109	0.200	0.124	0.197	0.211	0.293
Time[4]:PDGroup[2]	-0.096	0.201	-0.138	0.200	-0.132	0.295
Time[5]:PDGroup[2]	0.003	0.206	-0.040	0.206	-0.008	0.298

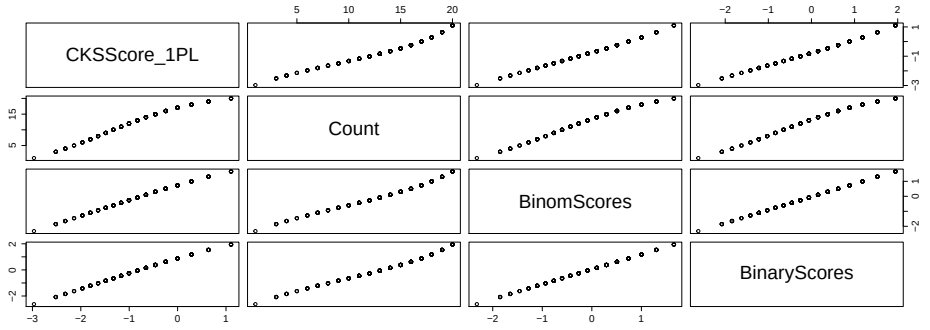


Figure 9: Scatterplots of scores from three different approaches with the original counts correct out of 20.

suggest little evidence of differences in coaching knowledge based on arms of the study or outside mathematics training but strong evidence of increasing scores over time and higher scores for coaches that had recent coach training outside of the project. The similarity of results is likely due to the simplicity of the initial scoring model and the limited number of items considered. With many more items and differential item discrimination, the mapping of the multivariate binary responses to scores might involve more than just finding the proportion correct. Here, though, it ends up that the three scoring approaches provided estimated scores that were perfectly related (Figure 9), with nonlinearity in the relationship between score scales and the counts providing the only noticeable differences. Similarly, all the analysis results were very similar. Some of the differences in p-values for effects is likely more due to testing methods than differences in strength of evidence - differences between F and χ^2 distributions would account for some differences in p-values. Table 3 provides the estimated model coefficients and standard errors and shows similar, although not identical, information identified by all three models for the effects of interest. If all the scores are related and analyses are similar, then the preference for methods might relate to simplicity of implementation or the additional information provided. For example, the binomial model provides no insight into item difficulties so is least likely to be favored.

The scoring and then analysis approach requires an independent data set, but provides some noticeable advantages. First, items can be assessed and potentially discarded without relying on the data to be used for the longitudinal data analyses. The next best option would be to use the initial observations to create a scoring model, but this suffers from a potentially small sample size and the researchers could be accused of screening items to provide clearer evidence of certain effects. Second, it provides scaling for the longitudinal responses that is interpretable relative to the variability and skill level of the subjects from the independent data set. In cases where a random sample from a population of interest is available, this interpretation is especially powerful. Third, linear mixed models are much easier to estimate and interpret than generalized linear mixed models, although the estimation challenges with GLMM have been decreased in recent years. Between the binomial and binary models, the computational complexity is greatly increased in attempting to model the binary responses but this method returns the benefit of information on item difficulty. That seems like a reasonable trade-off unless the binary-response model fails to converge.

In this situation, the differences between the independent data set and the longitudinal responses were less than might be encountered in other situations. The subjects in the longitudinal study started with lower abilities than the volunteer "experts" but the longitudinal scores of study participants increased over time and ended up approaching the mean of the

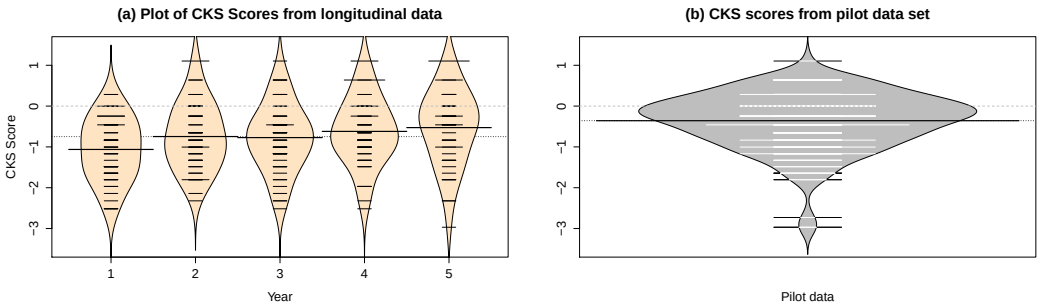


Figure 10: Bean-plot comparing the CKS scores from the longitudinal data (a) to those from the pilot data set (b).

volunteers by the end of the study. Figure 10 shows beanplots (Kampstra, 2008) of the scores from the pilot data set and the results for the subjects in the study by year, clearly showing the growth in expertise over time in the studied subjects. While we were not able to directly identify changes over time related to the specific coaching-related training the coaches received as part of the study, the identified impacts of being in the study or having outside training in coaching were to generally increase scores on the instrument. The positive changes over time in the study and impacts of outside coaching training suggest that the instrument is measuring what it claims - the knowledge of issues related to coaching. The similarity of the different analyses suggests that it is not an artifact of the independent data set or a particular modeling approach.

6. Acknowledgments

The authors would like to thank the research team and advisory board involved with the EMC project for feedback on the methods presented in this paper. Additionally, there would be no data to analyze without all the hard work of the teachers and coaches that participated in the project.

REFERENCES

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014), lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6.

de Ayala, R. (2008), *The Theory and Practice of Item Response Theory*, New York: Guilford Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., and Partchev, I. (2011), The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software* 39 (12), 1–28.

Doran, H., Bates, D., Bliese, P. and Dowling, M. (2007), Estimating the Multilevel Rasch Model: With the lme4 Package. *Journal of Statistical Software* 20 (2), 1–18.

Greenwood, M. and Banner, K. (2014), *A Second Semester Statistics Course with R*, scholarworks.montana.edu/xmlui/handle/1/2999

Kampstra, P. (2008), Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28 (1), 1–9.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and the R Development Core Team (2013), nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–113.

Reckase, M. (2009), *Multidimensional Item Response Theory*, New York: Springer.

Revelle, W. (2014), psych: Procedures for Personality and Psychological Research, Northwestern University.

Rizopoulos, D. (2006), ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17 (5), 1–25.

Sutton, J., Burroughs, E., and Yopp, D. (2011), Defining Coaching Knowledge. *Mathematics Education Leadership*, Fall, 12–20.