Challenges from Modeling Open Online Assessment Data

Yan Liu[1], Henrike Besche[2], Audrey Beliveau[3], Xingyu Zhang[4], Edward Kroc[1], Melanie Stefan[5], Johanna Gutlerner[2] & Chanmin Kim[6]

The University of British Columbia[1], Harvard University[2], University of Waterloo[3], The Hong Kong University of Science and Technology[4], Edinburgh Medical School[5] & Boston University[6]

Abstract

Open online assessment data is an emerging area of interest and has posed many modeling challenges for researchers. The existing psychometric models are not appropriate for these type of data because of several issues: (1) the data are usually comprised of small sample sizes; (2) a variety of topics are included in one short quiz, which make it impossible to use conventional multidimensional models; and (3) open online assessments bring in new information that have not been considered in the past, e.g., response time and response order data. This study explores the utilization of data visualization techniques, generalized linear mixed models, and latent class analysis for analyzing open online assessment data. Our study tries to fill in the gap between the advent of online classroom assessments and the lack of appropriate statistical models for analyzing these new types of assessment data.

Key Words: Open online assessment, response time, Bayesian generalized linear mixed model, data visualization, latent class analysis, psychometrics

Introduction

Over the past few years, open (or "open book") online assessments have gained popularity through Massive Open Online Courses (MOOCs). More and more high school teachers and university professors have started to use online assessments in their classroom. In recent years, open online quizzes have gained popularity in classroom assessment because they are easy to frequently implement, require little or no grading time, provide prompt feedback, and encourage higher cognitive thinking skills (e.g., Buchanan, 2000; Ibabe & Jauregizar, 2010; Miller, 2009; Rakes, 2008; Johnson & Kiviniemi, 2009). Open online assessments are typically low stakes and designed to give students flexibility to work on assessments inside or outside of the classroom. Students are able to review course materials, search for information, and develop a deeper understanding of the subject before finishing the assessments, a stark contrast to high-stakes closed book exams.

Open online quizzes are designed to guide and encourage student learning. However, online assessment data have posed many modeling challenges for researchers. Most existing psychometric models are not appropriate for modeling open classroom assessment data because of several issues. First of all, most psychometric models, such as item response theory (IRT), are developed for large scale assessment data and require a minimum sample size of 200 or 300; however, classroom sizes in most colleges or universities rarely exceed 100 or 200. As an example, van der Linden's hierarchical model (van der Linden, 2007) is a popular approach currently used by psychometrians to examine the relationship between response accuracy and response time. However, this approach

uses an IRT based model, which is not applicable to classroom assessment data with small sample sizes. In addition, the existing IRT based models are not flexible enough to accommodate many predictors in an analysis.

Second, many existing methods rely on an important assumption: uni-dimensionality. This assumption is often violated because one classroom assessment usually measures several topics. Some multi-dimensional models, such as multi-dimensional IRT, currently exist, but in a real open online classroom setting quizzes are usually short and sometimes test a topic with only a single question, making it impossible to apply multi-dimensional IRT.

In addition, the characteristics of open online assessment generate extra interesting information for researchers: for instance, item response time and the orders of taking items. With online assessments, the orders of taking items are available because an online system is able to record it. A quiz can be presented on one screen since it is usually short, which allows students to scroll back and forth and work on some easy questions first and then more difficult ones later. We have not found any research in the literature that considers response orders and attempt to model how response orders are related to student performance. Hence, there is a need for researchers to explore different statistical and psychometric methods to model open online assessment data.

Study Purpose

The purpose of the present study is to explore different statistical methods for modeling open book online assessment data and also to utilize data visualization tools to understand the characteristics of the data and identify relationships among variables. More specifically, using open assessment data collected from an undergraduate biology course in 2014 at Harvard Medical School as an example, we propose methods for addressing the following five research questions:

(i) How are item response times (i.e., time on each item) related to item responses (i.e., correct or wrong on each item) in terms of different levels of cognitive levels of the items?
(ii) Do students who passed the online quiz make more effort than those who failed?
(iii) What patterns can we find in the response orders when students answer questions?
(iv) How are item responses related to response times, the order of question attempts, learning strategies, and study time?
(v) Are there any latent groups (i.e., different response patterns) based on student actual responses to the quiz items?

We address (i) – (iii) via data visualization, (iv) using a generalized linear mixed model, and (v) using latent class analysis.

Contributions

The present study will explore how one can use information collected from open online quizzes to inform classroom instructors' teaching and student learning on a daily basis. This can help instructors see whether there is a need to improve classroom instruction. The present study will make contributions to the existing psychometric and statistical literature, filling in the gap between the advent of online classroom assessments and the lack of suitable psychometric and statistical methods for modeling this new type of data.

Methodology

Sample

A sample of 170 first year undergraduate students participated in this study. Students were enrolled in an undergraduate biology course at Harvard Medical School in 2014. The course instructors tried to use frequent online assessments to facilitate student learning for this introductory biology class, starting in 2013. The present study used data collected from the 2014 class.

Measures

Open online assessments

The assessments were designed for a five-week intensive course, consisting of 29 quizzes with a range of 5 to 15 multiple choice items. For the purpose of illustration, we only used one of the quizzes in the demonstrations. The data were collected using Learning Catalytics (a web-based learning platform), and included student responses and their response time on each item. Based on the revised Bloom's cognitive model (Krathwohl, 2002), the assessments included items at cognitive levels of factual knowledge (recall), comprehension (understanding), and application. The recall items were relatively easier than those at the levels of understanding and application.

These assessments included a unique feature: students were able to see the correct answer immediately after entering a response. To receive credits for a quiz, students were required to answer at least 50% of the items correctly. These circumstances may induce different motivational behaviors than assessments where answers are only available after the completion of a quiz.

A short survey

A short survey was also included in each online quiz. We used two questions from the survey. The first one asked students to identify all the learning strategies they had used among the following: attended lecture, reviewed lecture notes, watched lecture video, read textbook, joined study group, used others resources (e.g. web resources). The second question was a multiple-choice question where students were asked how much time they spent reviewing the course materials (none, up to an hour, 1-2 hrs, 2-4 hrs, or more than 4 hrs), The questions were added at the beginning of a quiz.

Variables

Dependent variables. Item response or item response correctness (0/1 for multiple choices)
Independent variables. Item response time (in minutes, which is transformed using base-10 log scale); orders of attempting quiz questions/response orders (quantified for each student using a classification method); cognitive levels of items (recall=0, understanding =1, application =2); survey Q-1 (learning strategies); and survey Q-2 (time reviewing course materials)

Data analysis

We explored three strategies for analyzing an open online assessment data set (one quiz): (a) data visualization; (b) generalized linear mixed modeling (GLMM) to model relationships of item responses, item response time, item-taking order and item cognitive levels; and (c) exploration of response patterns using latent class analysis (LCA).

Data visualization

We used the *ggplot2* R package for data visualization (Wickham, 2016). All visualizations are based on the raw data although response times were log transformed because of some extremely large values. Some students spent large amounts of time on individual quiz

questions. These outliers were handled differently in the GLMM analysis (see details in "missing values and outliers").

Generalized linear mixed model (GLMM)

To demonstrate how to model relationships of item responses and predictors, we used *Bayesian generalized linear mixed models (Bayesian GLMM)* with a logit link function. The *MCMCglmm* R package was used for the data analysis (Hadfield, 2010). Bayesian GLMM based on a Markov Chain Monte Carlo (MCMC) algorithm was used in our analyses. The iteration number was set to 50,000. The first 10,000 iterations were discarded as the burn-in period and the remaining 40,000 iterations were used for posterior computation. Results were reported for each parameter in the form of its posterior mean and corresponding 95% credible interval.

A brief description of Bayesian GLMM (Gelman et al., 2014; Zeger & Karim, 1991) is provided as follows. Let i denote item 1, …, n and j denote respondent 1, …, k and let $\mathbf{y_j} = (y_{1j}, \ldots, y_{nj})^T$ with y_{ij} be the response to item i by respondent j, then

$$E(y_j|\beta, u_j) = h(X_j\beta + Z_j u_j), \quad (1)$$

$$u_j \sim N(0, D) \quad (2)$$

where $h(\cdot)$ is a link function; β denotes the fixed effects (i.e., a vector of regression coefficients); u_j denotes the random effects (i.e., deviation score from the population mean of a parameter, such as the intercept or a slope); X_j and Z_j are the design matrices for the fixed effects and random effects, respectively. For β and D, we specify the following priors:

$$\beta \sim N(\tilde{\beta}, \tilde{\Sigma}) \quad (3)$$

$$D \sim W^{-1}(\Psi, \upsilon). \quad (4)$$

Note that the fixed effects β are assumed to have a multivariate normal distribution with a mean vector $\tilde{\beta}$ and a variance and covariance matrix $\tilde{\Sigma}$. The random effects are assumed to have a multivariate normal distribution with a mean of zero and a variance and covariance matrix D. D is assumed to have an inverse Wishart distribution with a scale matrix Ψ, which reflects the variation in the outcome variable across participants with degrees of freedom υ.

Latent class analysis (LCA)

The present study adopted latent class analysis (LCA) to explore whether there exist subpopulations with heterogeneous response patterns. LCA identifies subpopulations (i.e., latent classes), a type of finite mixture modelling with categorical indicators (Goodman, 2002; Asparouhov & Muthén, 2015; Vermunt, 2010). It also offers a wide range of fit statistics to help researchers evaluate the number of clusters. LCA is based on the assumption that there are one or more unobserved factors (i.e., latent variables) accounting for variation in observed variables and has a multinomial distribution (Collins & Lanza, 2010). The present study conducted LCA using the Mplus software program (Muthén & Muthén, 1998-2017), which is a structural equation modeling based program.

A brief description of LCA is provided as follows. A latent class analysis model for binary response variables to items $i = 1, 2, \ldots, I$ with C latent classes ($c = 1, \ldots, C$) can be expressed as

$$P(Y_j) = \sum_{c=1}^{C} \eta_c \prod_{i=1}^{I} \pi_{ic}^{y_{ij}} (1 - \pi_{ic})^{1-y_{ij}} \quad (5)$$

where η_c denotes the probability that an individual is a member of class c, y_{ij} denotes the observed binary response of individual j to item i; and π_{ic} denotes the probability of a positive response to item i from an individual from class c.

Missing values and outliers
Missingness is not a problem for this example data. Students took the quizzes seriously because they could earn credits if they were able to answer half of the quiz questions correctly. However, outliers on response times posed a challenge for statistical models. Students were given one day (during the weekdays) or 3 days (during weekends) to finish one quiz. The majority of students (79.4%) finished the quiz within a normal range, no more than 20 minutes per question, as estimated by the course instructors. For the other 20.6% of students, we did not know whether very large response times were due to spending time studying the course materials during the quiz or if they left their computer unattended for a long time. Hence, we treated outliers as missing values and conducted multiple imputation via the *mice* R package (Buuren & Groothuis-Oudshoorn, 2011) for the GLMM analysis.

Results

The results section is organized by three analytical strategies and five research questions. Data visualization results are reported to address the first three research questions, GLMM is reported to address the forth research question, and LCA is presented to address the last research question.

Data Visualization

For illustration purpose, we selected three graphics to demonstrate how assessment data can be visualized. Researchers have been trying to use different statistical methods for exploring whether response times (time used for answering each quiz/test question) can predict or inform student performance. Here we show that visualization is an important step in understanding open assessment data and assists one to see different aspects of the data, which are different from the information obtained from the statistical analyses.

RQ-1. How are item response times (i.e., time on each item) related to item responses (i.e., correct or wrong on each item) in terms of different cognitive levels of the items?
In this demonstration, we stratified the questions/items based on three cognitive levels, i.e., recall, understanding, and application. The course instructors designed the quiz questions based on the updated Bloom's Taxonomy (Krathwohl, 2002). It is assumed that recall questions require less cognitive ability (color coded in Figure 1) and should take less time (the median of response times on X-axis) and have higher item response correct rates (item easiness on Y-axis). On the other hand, quiz questions related to understanding or application skills may require more time and have relatively lower response correct rates. It should be noted that in classical test theory the percentage of getting a particular item correct is called item difficulty, which is denoted here as item easiness because the higher the correct rates, the easier the item is.

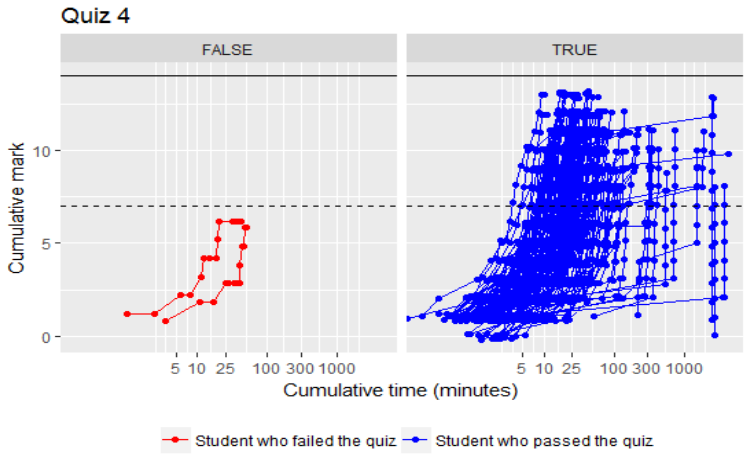
Figure 1 is used to address RQ-1 and shows that most items/questions are aligned with this assumption, but that items #13 and #14 need further investigation. Item #13 is a recall question, but less than half of students got this item correct, though students answered it in a reasonable time. Item #14 is an application question, and so is relatively challenging. Only 11% of students answered this item correctly, which perhaps suggests the question was too hard.

Figure 1. The relationships of item easiness, response times, and cognitive levels of items

RQ-2. Do students who passed the online quiz make more efforts than those who failed?

Figure 2 is used to compare students who passed the quiz with those who failed in terms of the time they spent on the quiz. The cumulative marks/grades are on the Y-axis, the cumulative time used for taking quiz questions is on the X-axis (log10 scale). Each line represents one student and each dot on the line denotes one quiz question. The figure shows that there is no obvious evidence that 168 students who passed the quiz made more efforts and spent more time (the median of total quiz times: 37.9 minutes) than the two students who failed (total quiz times: 40.9 minutes and 49.3 minutes). Because response time is on log10 scale, the slopes of lines seem steep and almost vertical for some students.

Figure 2. Comparison of students who passed the quiz with those who failed in terms of their response times (blue: student passed; red: student failed)

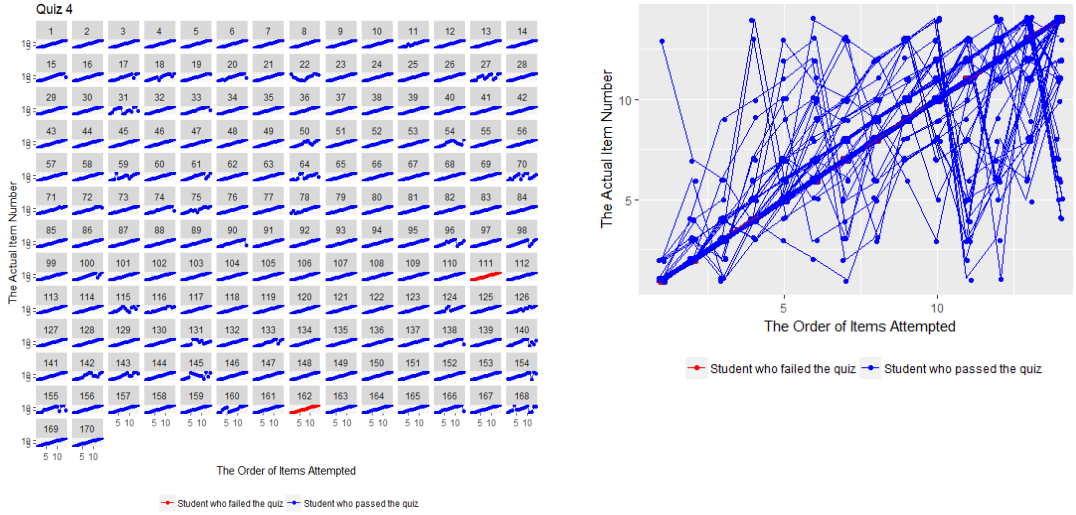


RQ-3. What pattern can we find from the orders of student attempting the questions?

Figure 3 is used to visualize the patterns in the orders of student quiz question attempts. The left graph is a small multiple plot for each individual student; the right graph shows

the patterns with all students in the picture. The Y-axis is the actual item number and X-axis represents the order of items attempted. For example, student #22 started item #13 (order #1), then moved to items #7 (order #2) and #6 (order #3), and so on. The majority of the students (65%) followed the linear order, starting with the first item and ending with the last item. However, some students jumped back and forth, which might be due to their test taking strategy or to their unfamiliarity with some course content. We did not interview these students, otherwise we could provide more explanations for the patterns we found here.

Figure 3. Orders of attempting quiz questions for each individual student (left) and for all students (right)



Generalized Linear Mixed Model (GLMM)

RQ-4. How are item responses related to response times, the order of question attempts, learning strategies, and study time?
We propose to use GLMM to address RQ-4, which allows one not only to examine the relationships between student responses and response times, but also to examine how orders of attempting the questions, learning strategies and study time affect student responses.

Table 1 presents the results of a GLMM analysis. We treated the data design as repeated measures (each student taking multiple measures/items) and set person to be a random effect in the analysis. It shows that student responses to the quiz items were related to cognitive levels of quiz items and the interaction of response time and recall quiz questions. The results suggest that, compared to application questions, students did better overall on recall questions (posterior mean of the cog.recall coefficient = 0.754, CrI =[0.481, 1.001], OR =2.126). Also, compared to application questions, students did worse overall on understanding questions (posterior mean of the cog.understd coefficient = -0.471, CrI =[-0.748, -0.190], OR =0.624). The interaction indicates that longer response time is associated with lower correct rate for recall items (posterior mean of the time*cog.recall coefficient = -0.081, CrI =[-0.155, -0.016]), but the interaction has a small effect size (OR =0.923). We did not find any statistically significant effect on responses from the other

predictors (i.e., the order of attempting quiz items, learning strategies and preparation time).

Table 1. Results of the GLMM analysis: Relationships of Item Responses and Item response Time, Orders of Item Taking, Learning Strategies, and Preparation Time

Fixed effects				
	posterior mean	odds ratio (OR)	95% CrI lower	upper
(Intercept)†	0.683	1.980	0.350	1.059
time	0.350	1.419	-0.004	0.751
order	-0.001	0.999	-0.005	0.002
lecture.review	0.229	1.258	-0.027	0.555
lecture.review.stdgroup	0.304	1.356	-0.145	0.696
lecture.more	0.168	1.183	-0.117	0.447
prep.time	0.061	1.063	-0.028	0.167
cog.recall†	**0.754**	**2.126**	**0.481**	**1.001**
cog.understd†	**-0.471**	**0.624**	**-0.748**	**-0.190**
time*cog.recall†	**-0.081**	**0.923**	**-0.155**	**-0.016**
time*cog.understd	-0.026	0.975	-0.093	0.048
Random effects				
person	0.019	1.019	0.001	0.055

† indicates variables for which the 95% CrI does not include 0.

Latent Class Analysis (LCA)

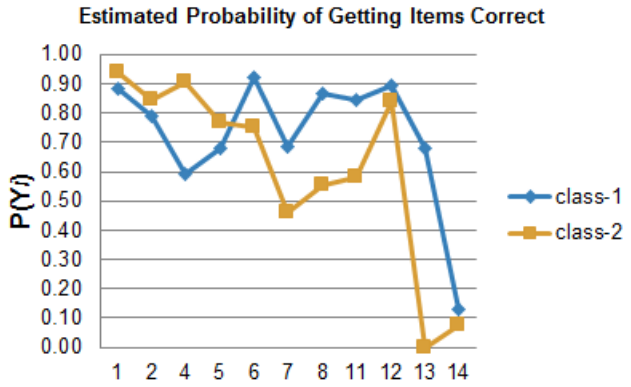
RQ-5. Are there any latent groups (i.e., different response patterns) based on student actual responses to the quiz questions?

Latent class analysis is used to explore whether there exist any different response patterns based on student actual responses to the quiz items. It should be noted that items 2, 9 and 10 were removed for this analysis because of their ceiling effects: about 100% of student answered these items correctly. A variety of fit statistics indicate the number of classes to retain, including the entropy value and four information-based fit indices: Akaike information criterion (AIC), Bayesian information criterion (BIC), sample-size adjusted BIC (SBIC) and Lo-Mendell-Rubin adjusted likelihood ratio test (LMR-A-LRT). For the information-based indices, a smaller value indicates a relatively better model fit. A p-value less than 0.05 in LMR-A-LRT indicates a statistically significant improvement from a model with (K-1) classes to a model with K classes. Two classes were finally decided based upon these criteria.

Figure 4 presents, for each of the two classes, the estimated probability of getting each item correct. The results suggest that students in class-1 made less efforts on the first three items, then they started to have relatively higher correct rates on items 6-12 compared to the other class. It is worth noting that both classes performed poorly on the last two quiz items (items #13 and #14), especially class-2. This result echoed what we found in Figure 1. Item #13 is not a hard question (it is a recall question), but none of the students in class-2 got this

item correct, whereas item #14 is an application question, but compared to other application questions a smaller number of students got this item correct (13% in class-1, and 7% in class-2). We would recommend that the instructors take a close look and explore potential explanations for these discrepancies. This may suggest some potential problems with the quiz question design or wording. In this demonstration, it may suggest that students made less efforts after they were assured that they answered at least half of the quiz questions correctly and were able to obtain the credits.

Figure 4. Results of latent class analysis with estimated probabilities of getting items correct by a line graph (left) and corresponding tabulated values (right)



Item	class-1	class-2
1	0.88	0.94
3	0.79	0.85
4	0.59	0.91
5	0.68	0.77
6	0.92	0.75
7	0.69	0.46
8	0.87	0.55
11	0.85	0.58
12	0.89	0.84
13	0.68	0.00
14	0.13	0.07

Discussion

The purpose of the present study was to fill in the gaps between advances in online classroom assessments and the lack of appropriate statistical models and visualizations for analyzing this new type of data. We explored three strategies for analyzing open book online assessments: data visualization, generalized linear mixed model (GLMM), and latent class analysis (LCA). We demonstrated that different methods could provide different information to classroom instructors and researchers.

Data visualization is a great tool for classroom assessment data. The sample size in a classroom is usually small, ranging from 10 to 200. Many existing psychometric or statistical methods are not applicable, especially when the sample has extremely unbalanced groups, such as 2 vs. 168 in Figure 2.

In addition, graphics can provide an efficient way of understanding our data. For example, Figure 1 shows a clear pattern of relationships of item easiness and median response times in terms of cognitive levels of items. It quickly helps one to identify two abnormal items, items #13 and #14, which are also identified by the LCA model.

Furthermore, graphics can show some information that statistical methods cannot achieve. For instance, Figure 3 shows the patterns in the orders of student quiz question attempts. One can see the pattern for each individual student as well as the overall pattern across all students. We assumed that students would start with relatively easy questions which they

were more confident in, and put off the harder questions until the end. Interviewing students could have provided useful information for classroom instructors to help understand which parts of the course content are relatively challenging to students and require more instruction or help.

Bayesian GLMM analysis is much more flexible than the existing IRT based models because it has less requirements on sample size and also allows the inclusion of multiple predictors. The results indicate that student responses to the quiz questions are related to item cognitive levels as well as the interaction of response times and item cognitive levels (recall) though the interaction has a small effect. In this demonstration, we quantified the orders of student quiz question attempts and generated one index for each student. However, this may not completely capture the relationship between individual quiz questions and the varying order of attempts for each student. We would encourage more studies to look into this issue.

LCA allows us to examine whether there exists a mixture of populations (latent groups). The LCA results suggest two latent groups with different response patterns. However, we did find both groups' performance on the last two quiz questions dramatically dropped, which is a warning to classroom instructors and indicates something may be going on. It is interesting to see that both data visualization and LCA revealed this information. Due to the particular design of the quizzes by the instructors, this dramatic dropping might be because students were not motivated after they were sure that they earned the credits from finishing the first 12 questions. In this demonstration, we only used student responses to quiz questions. Other researchers are encouraged to investigate response times using LCA if that is of interest.

There are currently not enough developed statistical or psychometric tools for analyzing classroom assessment data with small sample sizes, especially for the new format of open book online classroom assessments. The present study explored different strategies that are more suitable to class size data and provided some versatile tools to classroom instructors and researchers. We hope our study will motivate more research that explores diverse strategies for analyzing classroom assessment data and assists instructors to effectively guide student learning in the classroom.

References

Collins, L. M., & Lanza S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, And Health Sciences*. New York: John Wiley and Sons.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis, 1*, 515–534.

Goodman, L. A. (2002). Latent class analysis: The empirical study of latent types. In: Hagenaars JA and McCutcheon AL (eds) *Applied Latent Class Analysis.* Cambridge: Cambridge University Press, pp. 3–55.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The *MCMCglmm* R Package. *Journal of Statistical Software, 33(2)*, 1-22. URL http://www.jstatsoft.org/v33/i02/.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, 41(4), 212-218.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Eds. Los Angeles, CA: Muthén & Muthén.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45(3)*, 1-67. URL https://www.jstatsoft.org/v45/i03/.

Asparouhov, T., & Muthén, B. O. (2015). Residual Associations in Latent Class and Latent Transition Analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 22(2)*, 169-177, DOI: 10.1080/10705511.2014.935844

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika, 72*, 287. https://doi.org/10.1007/s11336-006-1478-z

Vermunt, J. (2010). Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis, 18*(4), 450-469. Retrieved from http://www.jstor.org/stable/25792024

Wickham, H. (2016). ggplot2: *Elegant graphics for data analysis*. Springer-Verlag New York,.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79−86.