frontiers in Education

ORIGINAL RESEARCH
published: 07 January 2021
doi: 10.3389/feduc.2020.607260

Bayesian Mixed Effects Model and Data Visualization for Understanding Item Response Time and Response Order in Open Online Assessment

Yan Liu[1]*, *Audrey Béliveau*[2], *Henrike Besche*[3], *Amery D. Wu*[1], *Xingyu Zhang*[4], *Melanie Stefan*[5,6], *Johanna Gutlerner*[3] and *Chanmin Kim*[7]

[1] Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada, [2] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada, [3] Harvard Medical School, Harvard University, Boston, MA, United States, [4] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, [5] Centre for Discovery Brain Sciences, The University of Edinburgh, Edinburgh, United Kingdom, [6] ZJU-UoE Institute, Zhejiang University, Haining, China, [7] Department of Statistics, Sungkyunkwan University, Seoul, South Korea

OPEN ACCESS

Edited by:
Jimmy de la Torre,
The University of Hong Kong,
Hong Kong

Reviewed by:
Ren Liu,
University of California, Merced,
United States
Zhehan Jiang,
University of Alabama, United States
Zhen Li,
eMetric LLC, United States

***Correspondence:**
Yan Liu
yan.liu@ubc.ca

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *16 September 2020*
Accepted: *30 November 2020*
Published: *07 January 2021*

Citation:
Liu Y, Béliveau A, Besche H, Wu AD, Zhang X, Stefan M, Gutlerner J and Kim C (2021) Bayesian Mixed Effects Model and Data Visualization for Understanding Item Response Time and Response Order in Open Online Assessment. Front. Educ. 5:607260. doi: 10.3389/feduc.2020.607260

Open (open-book) online assessment has become a great tool in higher education, which is frequently used for monitoring learning progress and teaching effectiveness. It has been gaining popularity because it is flexible to use and makes response behavior data available for researchers to study response processes. However, some challenges are encountered in analyzing these data, such as how to handle outlying response time, how to make use of the information from item response order, how item response time, response order and item scores are related, and how to help classroom teachers quickly check whether student responses are aligned with the design of the assessment. The purposes of this study are 3-fold: (1) to provide a solution for handling outlying response times due to the design of open online formative assessments (i.e., ample or unrestricted testing time), (2) to propose a new measure for investigating the item response order, and (3) to discuss two analytical approaches that are useful for studying response behaviors–data visualization and the Bayesian generalized linear mixed effects model (B-GLMM). An application of these two approaches is illustrated using open online quiz data. Our findings obtained from B-GLMM showed that item response order was related to item response time, but not to item scores; and item response time was related to item scores, but its effects were moderated by the cognitive level. Additionally, the findings from both B-GLMM and data visualization were consistent, which assisted instructors to see the alignment of student responses with the assessment design.

Keywords: open-book online assessment, open online assessment, classroom assessment, response time, response order, Bayesian generalized linear mixed effects model (B-GLMM), data visualization, Bloom's taxonomy

INTRODUCTION

Over the past decades, open (or "open book") online assessments have gained popularity through Massive Open Online Courses (MOOCs), such as Coursera (https://www.coursera.org/) and Edx (https://www.edx.org/). More grade school teachers and university instructors have started to use open online assessments as a learning and teaching tool. Among all types of formative assessments, open online quiz is probably the most widely used format because it is easy to implement, require little to no grading time, provide prompt evaluation and feedback to students and teachers (Buchanan, 2000; Brothen and Wambach, 2001; Rakes, 2008; Johnson and Kiviniemi, 2009; Ibabe and Jauregizar, 2010). In contrast to high-stakes, large scale, standardized achievement tests for ranking individuals (e.g., SAT and GRE), open online formative assessments are usually low-stakes, short and specific, designed to monitor, facilitate, and evaluate individual learning and teaching. Moreover, students can view the course materials while taking the test in a location and at a pace of their own wish.

Another advantage of a formative assessment, when computerized, is that it can automatically record a variety of information in addition to the students' answers and their marks. Response behaviors recorded in the open online assessment platform are called computer log data. Computer log data can keep track of student response behaviors in the process of completing the assessment. Such data can bring new and exciting insights that are difficult or impossible to obtain from a traditional in-class paper-and-pencil assessment. *Item response order* and *item response time* are two examples of such response behaviors. Additionally, the online platform can easily incorporate a survey along with the assessment, which can ask students to report their studying behaviors (e.g., time spent on studying) and learning strategies. Making good use of these data collected online not only can help monitor students' learning progress and provide tailored support in time, but also shed light on how student response behaviors are related to their assessment performance.

To date, reports on response behaviors are still scant in open online formative assessments compared to those in large scale standardized tests. As well, it is still a challenge as to how to quantify and analyze the response behaviors collected from open online formative assessments. This study tackles three data challenges arising from open online formative assessments: how to handle outlying response times, how to quantify item response order, and how these two response behaviors relate to the assessment outcome.

This study aims to study the relationships of item response time, response order and item scores in the open online assessment, which can be used to inform classroom teaching and learning. More specifically, the purposes of this study are 3-fold: to provide a solution to handle outlying response times, to propose a new measure for investigating item response order, as well as to showcase Bayesian generalized linear mixed effects model (B-GLMM) and data visualization as useful analytical approaches for studying response behavior data. The following paper is organized as follows: (i) discussing the issue of outlying item response times; (ii) proposing a new measure for quantifying item response order; (iii) discussing the flexibility of B-GLMM for studying item response behavior data; (iv) providing an illustration of B-GLMM and data visualization as useful tools for examining the relationships of items scores, item response time, item response order and other studying variables; and (v) providing a general discussion.

OUTLYING RESPONSE TIME

Rapid Guessing in Standard Testing

A preponderance of research on item response time has focused mainly on how to use the item response time to distinguish different response behaviors, rapid guessing vs. solution. Solution behaviors occurs when test takers actively answer the question and carefully provide an answer (Lee and Chen, 2011). In *high-stake* standardized testing, rapid guessing behaviors occurs mainly due to insufficient time, such that test takers rush their responses by the end (Schnipke and Scrams, 1997; Davey and Lee, 2011). In *low-stakes* testing, rapid guessing behavior occur because test takers were not motivated to try hard, such that test takers rapidly respond to items without making efforts (Klein and Hamilton, 1999; Wise and Kong, 2005; Wise and DeMars, 2006). Because rapid guessing behaviors can contaminate the item parameter estimation, a variety of psychometric methods have been developed to account for their effects (Schnipke and Scrams, 2002; Wise and DeMars, 2006; Klein Entink et al., 2008; van de Linden and Guo, 2008; Meyer, 2010).

Outlying Response Times in Open Online Formative Assessments

The challenge in open online formative assessments, however, is not rapid guessing, but outlying response times. Open online assessments usually give flexible and plenty of time for students to complete the test. As said earlier, the time flexibility is provided with a purpose to facilitate self-learning and self-evaluation rather than testing and ranking the examinees, so that students can view the course materials while taking the assessment. However, these flexibilities often result in outlying response times, i.e., a "unreasonably" long time was recorded to complete a question (e.g., hours or even days). Outlying response times can introduce noise to the time data if the reasons for it is unclear. It could be because the student was reviewing or studying the materials or because student left the quiz online while doing something else. If we know what causes the outlying response times, we can either adjust these response times accordingly or incorporate certain covariates to control for the confounding effects. However, the reasons for outlying response times usually are unrecorded and unknown.

Outlying response time is a commonly seen phenomenon in an open online formative assessment and need to be properly handled before data analysis. However, to our best knowledge, there are no reports regarding how to handle outlying response time. We propose to use a missing data imputation approach to handle the problem of outlying response time. This method will be described in the data analysis section of Demonstration-2.

ITEM RESPONSE ORDER

What Is Item Response Order?

Another essential information embedded in open online formative assessments is item response order, which is not available in a typical high-stake standardized test. *Item response order* refers to the order in which an individual chooses to answer the questions in a quiz or test. An open online quiz is usually

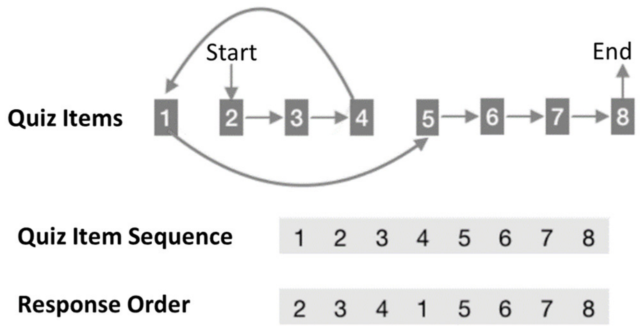

FIGURE 1 | An example of item response order comparing to the original quiz item sequence.

short, and the questions can be displayed on a single screen page. Students are able to scroll up and down to browse the questions anytime and do not have to follow the original *item sequence*, i.e., the sequence in which the items are presented. Taking a quiz with eight questions (see **Figure 1**), for example, a student can start from item #2, continue to items#3 and #4, before returning to item #1, and then jump to item #5 to complete the remaining questions. In this case, the student displays an item response order: 2-3-4-1-5-6-7-8. Such flexibility in item response order is usually not allowed in taking a standardized, computerized test where the items are, for the most part, ordered in a fixed form or delivered one by one adaptively, determined by some item selection algorithm if the test is adaptive. Item response order in a paper-and-pencil test is rarely studied because the response order is usually not recorded.

To our best knowledge, there is little to no research on item response order. The extant literature, however, has reported two kinds of phenomena that are related to, yet distinct from, item response order– item order effect and item position effect. *Item order effect* investigates how the same items that are ordered differently in two or several alternate test forms can affect the performance of examinees who are randomly assigned to the alternate forms (Monk and Stallings, 1970; Hambleton and Traub, 1974; Laffitte, 1984; Balch, 1989; Zwick, 1991). For example, Hambleton and Traub (1974) reported that students did better when the items were ordered from the easiest to hardiest. As for the *item position effect*, it focuses on the impact of the relative position of an item on its parameter estimates, such as equating parameter estimates in item response theory (Whitely and Dawis, 1976; Kingston and Dorans, 1984; Meyers et al., 2008; Li et al., 2012). For instance, Meyers et al. (2008) reported that the amount of change in the item position would affect the estimates of item difficulty in item response theory (IRT). These studies investigated how item order and item position in a standardized test affected the psychometric properties of test items and the students' performance, which are totally different phenomena from item response order that we investigated for open online assessment in the present study.

Because of lack of research on the item response order, it is still unknown as to what extent individuals' response orders deviate from the original item sequence and whether this deviation can provide insightful information about test takers' item response time and item performance. We conjecture that response order could be related to students' learning preparedness (e.g., familiarity with the course materials) and/or test-taking strategies. Yet, there is no existing measure for quantifying the characteristics of item response order so that its relationships with the other response behaviors and with the response outcomes can be examined quantitatively. Hence, we proposed a new measure to quantify item response order deviation as follows.

New Measure: Response Order Deviation (ROD) Measure

The ROD measure was developed with a purpose to summarize the extent to which a student's item response order deviates from the original item sequence. It is calculated for each individual using the following formula,

$$\mathrm{ROD} = \sum_{i=1}^{P-1} |R_{i+1} - R_i| + \sum_{i=1}^{P} |R_i - S_i| - (P-1) \qquad (1)$$

where P denotes the number of items ($i = 1, 2, \ldots, \mathrm{P}$), S_i stands for the sequence in which an item appears in the quiz, and R_i denotes the actual order in which an examinee responds to the *ith* item.

The first component of the equation (1), $\sum_{i=1}^{P-1} |R_{i+1} - R_i|$, indicates whether the response to an item R_i is followed by a response to the next item R_{i+1}. When a student did not follow the quiz item sequence, the first component indicates how far the student jumped from one item to another; the second component, $\sum_{i=1}^{P} |R_i - S_i|$, measures the distance between the response order number R_i and the sequence number of the original quiz item S_i for the *i*th item. If a student answers all questions in the original quiz item sequence, the first component will turn to $P - 1$ and the second component will be zero. The last component $-(P-1)$ is added so that the ROD $= 0$ when there is no deviation. The larger the ROD values, the further a student's response order deviates from the original item sequence. Note that the ROD measure can be standardized by P if one wants to compare the item response order across different assessments with different number of items.

Researchers can use the ROD measure directly as a quantitative variable in data analysis. Alternatively, the ROD measure can be categorized to an ordinal variable, such as low, medium, and high levels. In the section of empirical data illustration, we will demonstrate how individuals' ROD values function as an exploratory variable for item response time and item response outcome.

BAYESIAN GENERALIZED LINEAR MIXED EFFECTS MODEL (B-GLMM)

A variety of psychometric methods have been developed for modeling response behavior variables, especially examining the relationship of response time and response accuracy/latent ability. Unfortunately, many are only suitable for standardized summative tests and require a large sample size to achieve stable parameter estimation. For example, Wang and Hanson (2005) incorporated response time in the 3-parameter item response theory (IRT) model and treated response time as a fixed predictor to students' latent abilities. van de Linden (2007) proposed a hierarchical model to simultaneously analyze the relationships of response speed and accuracy with a combination of IRT and lognormal models using the multilevel approach. Klein Entink et al. (2008) extended van de Linden (2007) model by allowing the model to include predictors to explain the variance in speed and accuracy at the third level. Also with large-scale data in mind, several studies adopted IRT-based models to distinguish disruptive response behaviors (Schnipke and Scrams, 2002; Wise and Kong, 2005). For example, Meyer (2010) as well as Wise and DeMars (2006) utilized mixture IRT models to study rapid guessing behaviors in low-stakes large scale tests.

These IRT-based methods are not suitable for classroom assessments data because the size of the data is usually relatively small. An undergraduate class normally ranges from 100 to 300, which is too small for IRT, let alone the K-12 classrooms. We propose to use Bayesian generalized linear mixed effects model (B-GLMM) for modeling response behaviors. There are four advantages of using B-GLMM for analyzing open online formative assessment data. First, B-GLMM is a regression-based method and hence is less demanding on sample size than the latent variable based methods, such as IRT. Second, Bayesian approach allows one to assess the uncertainty in the parameter estimation, i.e., providing a probability distribution of the population parameter, rather than solely a point estimate. Third, it is recommended over the frequentist approach when the model is complex and the sample size is small (Browne and Draper, 2006; Gelman, 2006; Baldwin and Fellingham, 2013; Stegmueller, 2013). Additionally, B-GLMM is a fairly flexible analytical tool; It is easy to include explanatory variables and model them as having random or fixed effects. The B-GLMM can be seen as what De Boeck and Wilson (2004) called "explanatory item response models," but it extends the analytical models to a Bayesian approach.

The General Framework of B-GLMM

A general description of B-GLMM, based on Gelman et al. (2014), Wu (2009), and Zeger and Karim (1991), is provided as follows. Let i denote items $1, \ldots, I$ and j denote respondents $1, \ldots, J$ and let $\mathbf{y}_{\mathrm{j}} = (y_{1j}, \ldots, y_{nj})^{\mathrm{T}}$ where y_{ij} denotes the response to item i by respondent j, then

$$E\left(y_{ij}\middle|\beta, u_j\right) = h\left(X_{ij}\beta + Z_{ij}u_j\right), \tag{2}$$

$$u_j|D \sim N(0, D) \tag{3}$$

where $h(\cdot)$ is a link function (e.g., identity, logit, or log); β denotes the fixed effects (i.e., a vector of regression coefficients); u_j denotes the random effects (i.e., the deviation scores from the population mean of a parameter such as the intercept or a slope), which are assumed to have a multivariate normal prior distribution with a mean of zero and a variance and covariance matrix D; X_j and Z_j are the design matrices for the variable having fixed effects and variables having random effects, respectively. For β and D, we define the following priors:

$$\beta \sim N\left(\tilde{\beta}, \tilde{\Sigma}\right) \tag{4}$$

$$D \sim W^{-1}(\Psi, \upsilon). \tag{5}$$

Note that the fixed effects β are assumed to have a multivariate normal prior distribution with a mean vector $\tilde{\beta}$ (usually a set of zeros) and a variance and covariance matrix $\tilde{\Sigma}$ where the covariances are usually set to zeros and variances set to some large values that are uninformative. The D is assumed to have an inverse Wishart distribution denoted by W^{-1}, which reflects the variation in the outcome variable across individual subjects with a scale matrix of Ψ and degrees of freedom υ.

The Bayesian posterior distribution of all parameters given the data is proportional to the product of the likelihood and the prior distributions:

$$f\left(\beta, D, u\middle|y\right) \propto \left(\prod_{i=1}^{I}\prod_{j=1}^{J} f\left(y_{ij}\middle|\beta, u_j\right)\right)\left(\prod_{j=1}^{J} f\left(u_j\middle|D\right)\right) f(\beta) f(D). \tag{6}$$

The posterior distribution represents the updated belief about the parameter, after considering the observed data under a given model. B-GLMM based on Markov Chain Monte Carlo (MCMC) algorithm is used for the present study. In the following two sub-sections, we describe how B-GLMM can be used for modeling item response time and item performance. These two B-GLMMs will be demonstrated in the section of Real Data Illustration.

Modeling Item Response Time

As we described above, B-GLMM is a flexible modeling tool, which can incorporate explanatory variables and allow them to be either having random or fixed effects. Generally, researchers can collect explanatory variables, such as students' prior knowledge, learning strategies, and course preparation, that are related to their academic performance. In this study, we focus on the explanatory variables that can help uncover student cognitive processes and test-taking behaviors during the assessment. Two essential explanatory variables, item response order and item cognitive level, are included to model item response time. Each is explained as follows.

Item response order becomes available to researchers as open online assessments allow the students to respond to test questions in different orders. Because there is yet research on item response order, it is still unknown as to what extent the individuals' response orders will deviate from the original item sequence and whether this deviation can provide insightful information about students' item response time and item performance. These questions will be investigated using the ROD measure we proposed earlier in this paper.

Additionally, we investigated how item cognitive level affected response time. Items in a formative assessment are usually developed based on some learning or pedagogy theory related to cognitive development. For instance, Bloom's taxonomy is regarded as the backbone of teaching and learning in K-12 and higher education (Biggs and Tang, 2011). The revised Bloom's taxonomy is a hierarchy of cognitive levels, from simple to complex and from concrete to abstract, including *Recall, Understand, Apply, Analyze*, *Evaluate*, and *Create* (Anderson and Krathwohl, 2001; Krathwohl, 2002). In order to assess student cognitive level of mastery, each of the items in a test is purposefully designed to target a certain cognitive level. Some items, for example, will only require the Recall level but others will require the Apply level to answer the item correctly.

We hypothesize that the item cognitive level will influence item response time. Items that require higher cognitive levels to answer correctly will take more time for students to answer. Additionally, we hypothesize that item cognitive level may moderate the relationship between item response time and item performance, which is described in the next section. We did not find any literature reporting how it affects response time and performance despite that item cognitive levels are frequently built into test questions.

Because item response time is a continuous variable, the identity link function $h(\cdot)$ in Equation (2) is used for B-GLMM. Also, it is very likely that the response time data is skewed; in this case, it is advised to transform the data using natural logarithm. The B-GLMM for response time is specified as follows,

$$Log(Time_{ij}) = \beta_0 + \beta_1 ROD_j + \beta_2 Cog_i + \beta_3 OtherCov_j + u_j + r_{ij} \quad (7)$$

where $i = 1, \ldots, I$ denotes items and $j = 1, \ldots, J$ denotes respondents; the outcome is the log of response time (Time). As suggested by Wilson et al. (2008), the items are treated as the repeated observations and clustered under individual subjects, which is a multivariate analysis approach. The predictors ROD, item cognitive level (Cog), and other learning-related covariates (OtherCov; see the Real Data Illustration) are treated as having fixed effects. Individuals were treated having random effects u_j (person effect) on response time. The r_{ij} are the residuals, which we take as normally distributed with equal variances.

Modeling Item Performance by Item Response Time and Order

In this B-GLMM, item scores are included as the outcome; item response time, ROD, item cognitive level (Cog), and other covariates (OtherCov) are included to examine how they predict students' item performance. Additionally, we hypothesized that item cognitive level would moderate the effect of the item response time on item scores. Because the outcomes are binary categorical variables, the logit link function is used for this analysis. The model is specified as follows,

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Cog_i + \beta_3 Time_{ij} {*} Cog_i + \beta_4 ROD_j + \beta_5 OtherCov_j + u_j \quad (8)$$

$$y_{ij} \sim Bernoulli\left(p_{ij}\right) \quad (9)$$

where outcome y_{ij} denotes the item score (0 vs. 1) of item i by respondent j; p_{ij} denotes the estimated probability of each student succeeding for each item; all the other notations are the same as in Equation (7). As before, individuals are specified to have random effects (u_j) on item performance, representing individuals' differences (person effect) in understanding the topics covered in the quiz. All of the predictors, including their interaction terms, were treated as having fixed effects.

REAL DATA ILLUSTRATION

Using a real open online quiz data, this section is to showcase two analytical approaches for analyzing response behavior data: Data visualization and B-GLMM. We demonstrate how these methods can help researchers and educators to understand students' response processes through item response time, item response order, and item cognitive level and their effects on item performance. The findings may reveal insightful information for understanding teaching and learning. We first describe the sample data and then provide two demonstrations. By addressing five specific research questions, we hope to make the demonstrations more relevant to researchers and educators' practices while we showcase these analytical strategies.

Information About the Data

Sample

A sample of 170 first year undergraduate students participated in this study. Students were enrolled in an undergraduate biology course at a Medical School in the east coast of the U.S. in 2014. The course instructors have used frequent online assessments to monitor and facilitate student learning for this introductory biology class.

Measures

Open online assessment

The assessments were designed for a 5 week intensive biology course, consisting of 29 quizzes. For the purpose of illustration, we only used one of the quizzes. The data were collected using Learning Catalytics (a web-based learning platform) that included students' answers, response times and response orders. Based on the revised Bloom's cognitive model (Krathwohl, 2002), the instructors categorized each of the 14 multiple-choice items to one of the three levels of cognitive processes: factual knowledge ($Recall = 0$), comprehension ($Understand = 1$), and application ($Apply = 2$). The *Recall* items were relatively easier than the *Understand* and *Apply items*, and the *Understand* and *Apply* items could be equally challenging to students.

Learning behaviors survey

A short survey was also administered at the beginning of the online quiz to collect the information about students' learning behaviors. This study used two questions from the survey. The first question asked students to identify all the learning strategies they had used from a pre-specified list: (a) attended lectures only; (b) watched lectures and reviewed materials in addition to attending lectures; and (c) using more resources than just reviewing lectures and materials (e.g., web resources). The second question asked how much time students spent on reviewing the course materials before taking the quiz (none, up to an hour, 1–2, 2–4 h, or more than 4 h).

Variables

Outcome variables

Two outcome variables were investigated in the B-GLMMs, respectively. The first was the natural logarithm of *item response times* (in minutes) of the 14 multiple choice items, and the second was *item scores* (1 = correct, 0 = incorrect).

Explanatory variables

The two explanatory variables taken from the survey were *learning strategies* and *study time*. *Learning strategies* was considered as a nominal categorical variable, and hence two dummy variables were created with "attended lectures only" as the reference category. *Study time* was treated as a quantitative variable (none = 0, up to 1 h = 1, 1 to 2 h = 2, 2 to 4 h = 3, more than 4 h = 4). The variable *item cognitive level* (Recall = 0, Understand = 1, Apply = 2) was used as a categorical covariate for modeling item response time and was treated as a moderator for modeling item scores, moderating the effect of item response time on item performance. The measure of response order deviation (ROD) introduced earlier was used to model both outcome variables.

The relationships between these explanatory variables and the two outcome variables are explored first using data visualization techniques in Demonstration-1 and then using B-GLMM in Demontration-2. Each demonstration consists of three sections: research questions, data analysis, and results.

Demonstration-1: Data Visualization

Researchers have been using complex parametric psychometric models to understand the relationship between item response time and item scores. Data visualization is not regarded as a mainstream analytical tool in the literature. However, it can become extremely useful for understanding and exploring some unknown characteristics and relationships of open online assessments. The relational patterns among our variables are mostly unknown, such as the relationships of item response time, item response order and item scores. Hence, we start from data visualization method because it is a great exploratory tool and also can inform the statistical analyses in Demonstration-2. Specifically, data visualization was used to address the following research questions:

RQ-1: How are the item response time and item easiness (i.e., the correct answer rate) related, broken down by Bloom's cognitive levels (i.e., Recall, Understand, and Apply)?

RQ-2: What patterns can we find from students' item response orders?

RQ-3: How does item response order affect item response time and item scores?

Data Analysis

We used the *ggplot2* (Wickham, 2009) and *plotly* (Sievert, 2020) *R* packages for all data visualization (see **Figures 2–4**). The raw data were used for all visualization except for **Figure 4**, in which response times were log10 transformed to handle the outliers. The interactive graphics of **Figures 3B**, **4** can also be found on *Rpubs* (https://rpubs.com/yanliu/ItemResponseOrder and https://rpubs.com/yanliu/ResponseBehaviors).

Results

Results for RQ-1

It was expected that the *Recall* questions would require less cognitive ability and should take less time and tend to have higher correct rates (i.e., easier), whereas *Understand* and *Apply* questions would require more time and tend to have relatively lower correct rates. **Figures 2A,B** allowed educators to check the alignment of student responses with the quiz design. **Figure 2A** shows the relationships of item easiness (item correct rates), item response time and the item cognitive level. **Figure 2B** helps instructors gain a clear idea about the distribution of the item response time for each item[1]. Items # 9 and #10 stood out as the easiest items with less time requirements in **Figure 2A**, which were also shown to be more skewed with less variation in **Figure 2B**. Item #3 was a recall item, but it appeared to require more understanding and more time. The course instructor confirmed that these were expected because most students had learned related topics for items #9 and #10, which led to higher correct rates and less response times, whereas item #3 was a new topic and might require more time to recall a large amount of reading materials even if it was designed as a low cognitive level item.

The relationships among item cognitive levels, the medians of item response time (in minute), and item easiness (the proportion of correct answers) were consistent with our conjectures for most items, except items #13 and #14 that needed further scrutiny. Item #13 was a *Recall* question, but only less than half of students answered it correctly, despite that the students answered it in a reasonable time range. Item #14 was an *Application* question and should be relatively more challenging, but only 11% of students answered this item correctly. The course instructor examined these two items and indicated that item #13 might be confusing to students who did not have a deeper understanding on that topic, which could be improved, and item #14 was too hard to this student population, which suggested that there was a need to improve the instruction or add more exercise.

It should be noted that item easiness values were very high for items #9 (97%) and #10 (98%), leading to very low score variation. When the item score variation is low (almost everyone answered the item correctly), the data provide little or no information. This can complicate the estimation of a parametric method (e.g., IRT), leading to model convergence problems where one needs to remove these items to achieve model convergence). In contrast, data visualization works well in this scenario. We were able to examine the relationships of among item easiness, response time, and item cognitive level despite that some items had almost no variation.

[1] Given that 2.35% extreme outlying response times greatly distorted the distribution, we replaced these outliers by imputed values instead in the histogram plots. We could use log response time, but it would be hard for educators who do not have statistics background to understand the log time.

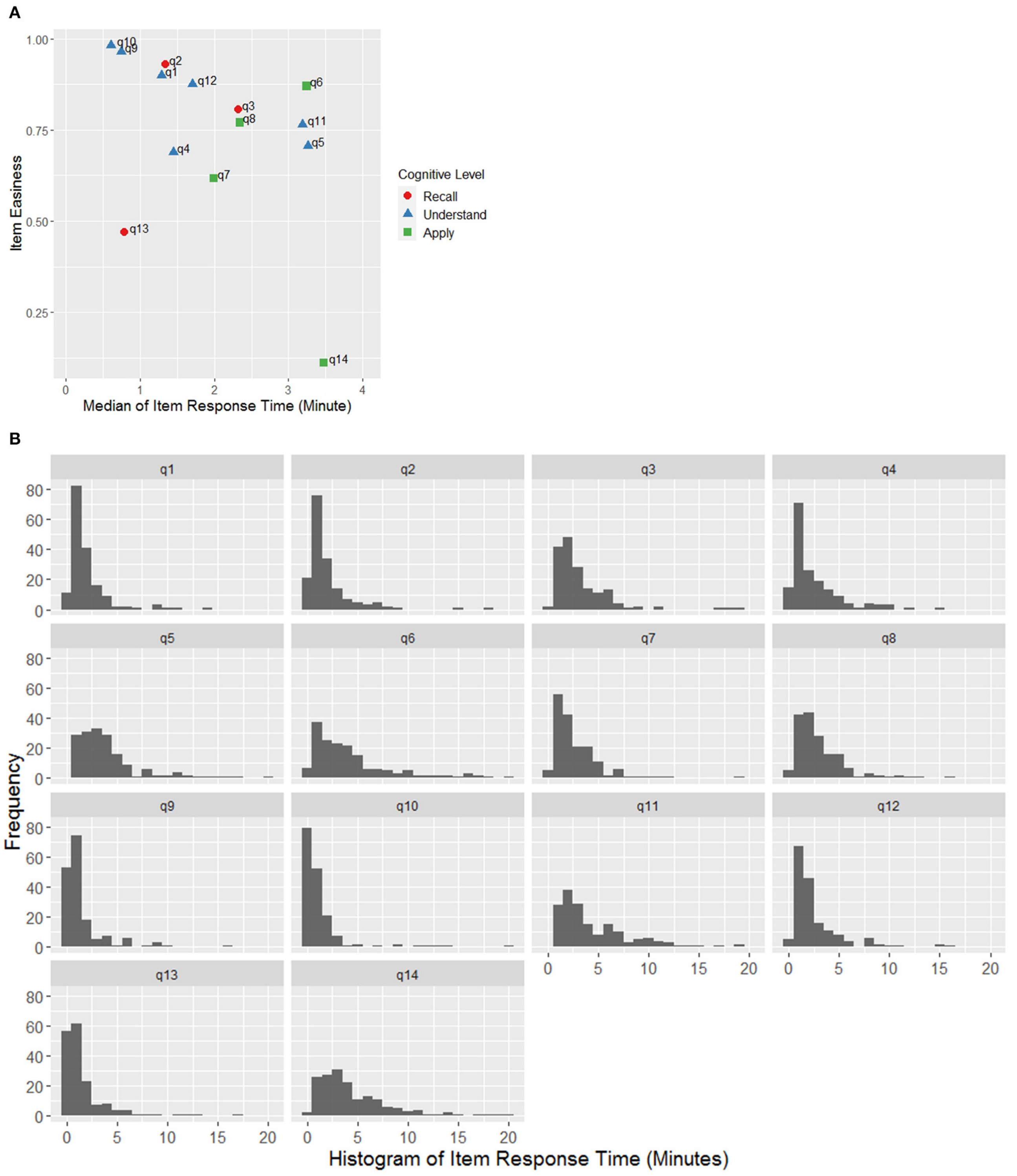

FIGURE 2 | Alignment of student responses with the quiz design. **(A)** The relationships among item easiness (the proportion of correct answers), the median of item response time (minute), and item cognitive level (Recall, Understand, and Apply). **(B)** Distribution of item response time for 14 Items (minutes).

Results for RQ-2

As we discussed in the previous section, item response order in open online assessments has not been researched in the literature. Data visualization can be a great tool used to explore some unknown characteristics of item response behavior data. **Figure 3** shows the patterns of item response order. Each cell in **Figure 3A**

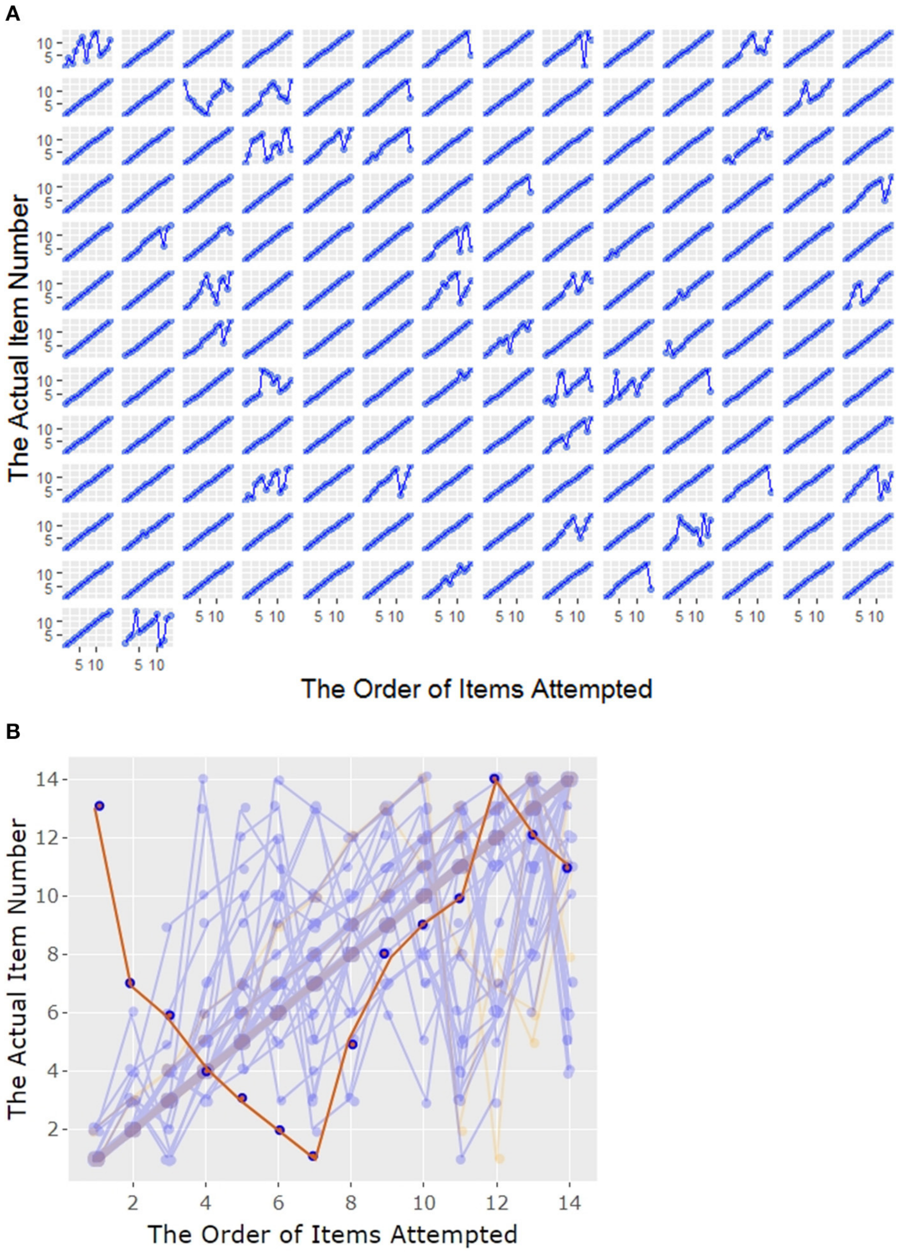

FIGURE 3 | Item response orders for individual students. **(A)** Small multiple plots: Item response order for each student. **(B)** Item response orders for all students in one graph with the subject #22 highlighted in red.

shows the item response order of a student. **Figure 3B** shows the item response order for all students in one graph. The Y-axis is the original quiz item sequence (from items #1 to #14) and the X-axis represents individuals' item response orders.

Taking subject #22 for example (row-2 and column-3 in **Figure 3A**, which is the line highlighted in red in **Figure 3B**), this student started with item #13, then moved to items #7 and #6, and so on. **Figure 3** also shows that the majority of the students (about 74%) followed the original quiz item sequence, as indicated by linear diagonal lines. However, some students jumped back and forth resulting in zigzag lines, which could be due to their test taking strategies or their unfamiliarity with the course content. A post quiz interview would be helpful to find out why these students chose their particular sequence to answer the questions.

Results for RQ-3

It is still unknown whether and how the item response order is related to item response time and item scores. **Figure 4** is used to address this question. Students' cumulative time used for taking the quiz is on the X-axis (shown on the log10 scale), and their

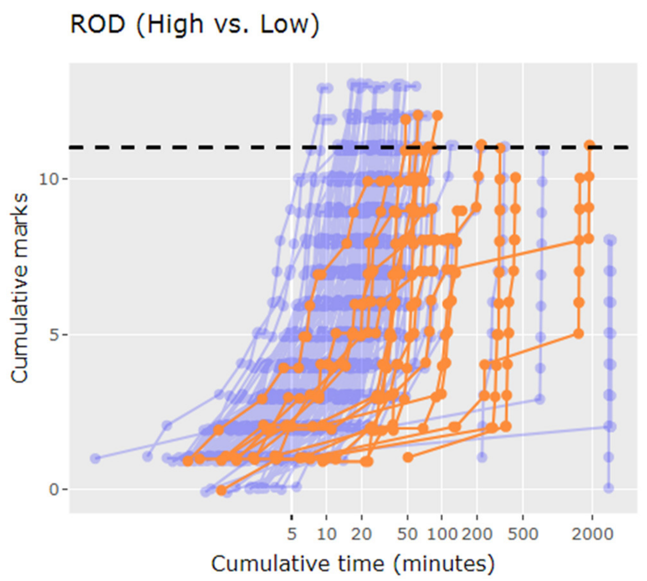

FIGURE 4 | Comparing response times and item scores between those who strictly followed the quiz item sequence (blue) and those who highly deviated from the quiz item sequence (orange).

cumulative mark is shown on the Y-axis. Each line represents one student and each dot on the line indicates one quiz item. The lines in light blue represent students who strictly followed the original item sequence (i.e., ROD values = 0), and the lines in orange highlight students who deviated highly from the original item sequence (i.e., ROD values > 50)[2]. It should be noted that the slopes of lines are fairly steep and almost vertical for some students because item response time is on the log10 scale.

The overall pattern shows that students who had a high level of ROD due to jumping back and forth used much more time to answer each item. Given that response time is shown on the log10 minute scale, the differences in response time between the two groups were large. Most of the students with high ROD scored lower than the 80% threshold (dashed line in **Figure 4**), in sharp contrast to those with ROD = 0 who scored mostly above the threshold. However, we did not know whether the observed difference was statistically significant. This will be further examined in Demonstration-2.

Demonstration-2: Bayesian Generalized Linear Mixed Effects Model (B-GLMM)

In this demonstration, we focused on the generalized linear mixed effects model using a Bayesian approach to investigate the relationships among item response time, item response order, item cognitive levels, and item scores. This statistical approach was used to confirm the findings obtained from the data visualization. The low data variation issue is always a challenge for any statistical methods. We compared the models with and without the two items that had very high correct rates of 97% and 98% and did not find them causing any modeling issues and the results were almost the same, so we included these two items in all the B-GLMMs. Specifically, it addressed research questions RQ-4 and RQ-5 below:

> ***RQ-4***: How did item cognitive level, item response order, learning strategies, and study time affect item response time?
> ***RQ-5***: How did item response time, item response order, learning strategies, and study time affect item scores? Did item cognitive level moderate the effect of the item response time on item scores?

Data Analysis

B-GLMM

As described earlier, two B-GLMMs were conducted for modeling item response time (Equation 7) as well as modeling item performance (Equation 8). The *MCMCglmm R* package (Hadfield, 2010) was used for the analyses. This *R* package is well-suited for performing B-GLMM, but it only allows users to run one chain by default. In order to check model convergence, we used the *parallel R* package (R-core, 2018) and ran 3 chains simultaneously.

The iteration number for the Markov Chain Monte Carlo (MCMC) algorithm 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 distribution and the computation of the summary statistics (e.g., posterior mean). Non-informative priors were used for this study. The fixed effects β were specified to have a multivariate normal distribution with a mean vector $\tilde{\beta} = 0$, variances of 10^8, zeros for covariances (see Equation 4). Both residuals r_{ij} and the random effect u_j (i.e., person effect) from Equations 7 and 8 were specified as a univariate normal distribution with a mean of zero and the prior variance followed an inverse Wishart distribution with a scale factor $\Psi = 1$ and degrees of freedom $\upsilon = 0.002$.

The model convergence was examined by Gelman-Rubin diagnostic (Gelman and Rubin, 1992). Gelman-Rubin diagnostic is a very useful method to examine model convergence. It compares both within- and between-chain variability. Gelman and Rubin (1992) and Brooks and Gelman (1998) suggested that the values of diagnostic <1.2 indicate convergence. The *gelman.diag*() function from the *coda* R package was used for obtaining Gelman-Rubin diagnostic (Plummer et al., 2006).

Results were summarized for each parameter estimate using the posterior mean estimate and its 95% credible interval (95% CrI). Odds ratio (OR) was used when the logit link function was applied. OR is a measure of effects size. As recommended by Cohen (1988) for OR, a small effect = 1.5, a medium effect = 3.5, and a large effect = 9.

The 3-step procedure for treating outlying response times

The outlying response times poses a challenge to data analysis because the time length recorded by the computer is not the

[2] The continuous ROD measure was categorized into three levels in order to contrast the low and high levels of ROD: (1) low deviation: ROD = 0 (74%), (2) medium deviation: 0 < ROD < 50 (16.5%), and (3) high deviation: ROD ≥ 50 (9.5%). The cut offs were chosen based on the distribution of the ROD values. We decided to use a cut off of 50 because a small proportion of students with ROD <50 behaved markedly different from the rest.

actual time used to answer a question. For the quiz used in this study, students were given up to 2 days to finish one short quiz. The majority of students (79.4%) finished the quiz within a normal time range, with no more than 20 min per item, as estimated by the course instructors. For the other 20.6% of students, we did not know what happened that led to outlying response time.

The outlying time data points were handled using a 3-step procedure. First, researchers need to decide on a cut-off score for defining whether the recorded response time was an outlier. We used 20 min based on the instructor's estimate on the time limit for students to answer the quiz questions, which was also confirmed by our examination of the distribution of item response time for each item. If a student used more than 20 min to answer a question, that recorded time was removed and treated as missing data.

In the second step, researchers need to conduct a missing imputation analysis to impute these outlying response times. Multiple imputation method is recommended and normally 5–10 data sets imputed can help to achieve a steady estimate of the missing value. In this demonstration, we conducted multiple imputations using *mice R* package assuming missing at random (van Buuren and Groothuis-Oudshoorn, 2011) and obtained 10 imputed data sets. In the missing data imputation, we included all the item scores, item response time, ROD, all survey questions obtained for the present quiz as well as the total scores obtained from the other four quizzes.

In the last step, researchers need to conduct analysis for each imputed data and then provide one integrated result. In the demonstration, we conducted B-GLMM analyses for each imputed dataset first and then combined the results for the posterior means and credible intervals for all the parameters as well as the odds ratios if the logit function is used (Gelman et al., 2014; Zhao and Long, 2017).

TABLE 1 | Results of B-GLMM for modeling item response time by item cognitive level, item response order, learning strategies, and study time.

	Posterior mean (log)	95% CrI (log)	
Fixed effects		**Lower**	**Upper**
(Intercept)†	0.44	0.17	0.71
lecture.review	−0.17	−0.44	0.10
lecture.review.more	−0.14	−0.41	0.14
prep.time	−0.01	−0.10	0.07
ROD.medium	**0.26**	**0.06**	**0.46**
ROD.high†	**0.53**	**0.28**	**0.79**
cog.understand	**0.12**	**0.03**	**0.21**
cog.apply†	**0.60**	**0.50**	**0.70**
Random effects	**Posterior mean of variance**	**95% CrI**	
		Lower	**Upper**
Person effect	0.18	0.14	0.24
Residual	0.72	0.68	0.76

CrI, credible interval; parameter estimates in bold † *CrI of a variable did not include 0. The reference group for the "learning strategies" variable was the group attended lectures only, against which the other two groups were contrasted: "lecture.review" (reviewed lectures and textbooks in addition to attending lectures), and "lecture.review.more" (joined study group and used other resources additionally). Meanwhile, the variable "prep.time" stands for the survey variable "study time"; the variable "time" refers to "item response time." "ROD.medium" denotes the medium level of response order deviation scores (0 < ROD < 50), and "ROD.high" denotes the high level of response order deviation scores (ROD > 50). Also, the reference category for the variable "item cognitive level" was "Recall," against which the two other levels were compared: the "cog.understand" and the "cog.apply".*

Results

Results for RQ-4

The results of Gelman diagnostic statistics ranged from 0.999 to 1.02, which were close to one, a criterion for excellent model convergence. **Table 1** presents the results of B-GLMM for modeling item response time. The 95% CrIs of posterior means of four variables (two ROD variables and two item cognitive level variables) did not include zero, ROD.medium (coefficient = 0.26, CrI = [0.06, 0.46]), ROD.high (coefficient = 0.53, CrI = [0.28, 0.79]), cog.understand (coefficient = 0.12, CrI = [0.03, 0.21]), and cog.apply (coefficient = 0.6 (CrI = [0.49, 0.72). The results suggested that item response order and item cognitive level were related to the item response time. The higher deviation from the original item sequence led to more response time. The items with higher levels of cognitive ability required more time to answer. The findings echoed what we found in **Figures 2**, **4** in Demonstration-1 by data visualization.

Results for QR-5

The poorer performance of students with high ROD scores observed from data visualization is tested here. Also, we hypothesized that item response time would affect item scores and such effect was moderated by item cognitive level. The response time variable was mean centered to prevent multicollinearity with the interaction terms for studying the moderation effects.

The results of Gelman diagnostic statistics also showed an excellent model convergence, Gelman diagnostic statistics ranged from 0.994 to 1.178. **Table 2** presents the results of the B-GLMM for modeling item performance. The 95% credible intervals (CrI) showed that students' item scores were related to item response time, item cognitive levels, and the interaction of these two variables. However, they were not found to be related to ROD and other learning related covariates. More specifically, three main effects were found to have small effects: response time (coefficient = 0.85, CrI = [0.60, 1.12], OR = 2.35), cog.understand (coefficient = 0.48, CrI = [0.14, 0.81], OR = 1.61), and cog.apply (coefficient = −0.93, CrI = [−1.28, −0.60], OR = 0.39).

Two interactions were found to have medium effect sizes: time*cog.understand (coefficient = −1.50, CrI = [−1.83, −1.19], OR = 0.22), and time*cog.apply (coefficient = −1.26, CrI = [-1.60, −0.92], OR = 0.28). For the ease of interpretation, the reciprocals of ORs were taken when the regression coefficients were negative. The ORs of these two interaction terms were 4.5 (1/0.22) and 3.52 (1/0.28). The results indicated that the relationship of response time and item scores was dependent on the cognitive level of the item. More specifically, a longer

TABLE 2 | Results of B-GLMM for modeling item performance by item response time, item response order, learning strategies, and study time.

Fixed effects	Posterior	95% CrI (logit)		Odds Ratio
	mean (logit)	**Lower**	**Upper**	**(OR)**
(Intercept)†	1.14	0.65	1.61	3.11
lecture.review	0.26	−0.17	0.68	1.29
lecture.review.more	0.17	−0.26	0.60	1.18
prep.time	0.12	−0.02	0.26	1.13
ROD.medium	−0.04	−0.36	0.29	0.96
ROD.high	0.16	−0.26	0.58	1.17
response time†	**0.85**	**0.60**	**1.12**	**2.35**
cog.understand†	**0.48**	**0.14**	**0.81**	**1.61**
cog.apply†	**–0.93**	**–1.28**	**–0.60**	**0.39 (2.54)**
time*cog.understand†	**–1.50**	**–1.83**	**–1.19**	**0.22 (4.50)**
time*cog.apply†	**–1.26**	**–1.60**	**–0.92**	**0.28 (3.52)**
Random effects	**Posterior mean of**	**95% CrI**		
	Variance	**Lower**	**Upper**	
Person effect	0.03	0.01	1.11	

The abbreviations for explanatory variables are the same as ***Table 1****. CrI denotes credible interval; parameter estimates in bold † denotes that the CrI of a variable did not include 0. The reciprocals of ORs were also provided for two interaction terms and "cog.apply" that had negative regression coefficients.*

response time was associated with a lower correct rate for both *Understand* (vs. Recall) and *Apply* (vs. Recall) items.

GENERAL DISCUSSION

With a goal to inform teaching and learning in the classroom, the purpose and design of open online formative assessments is quite different from a large-scale standardized summative assessment. Data collected from open online formative assessments can display very different characteristics, such as outlying response time and irregular item response order. We discussed these two essential issues arising from open online assessments and provided potential solutions, that is, multiple imputation for handling outlying item response time and the new measure ROD for studying the effects of item response order.

In the context of addressing five substantive questions, we adopted two analytical approaches, data visualization and B-GLMM, that are useful for modeling response behaviors as well as informing learning and instruction, and assessment development. The major findings obtained from our data visualization were supported by the B-GLMM analyses. Two findings obtained from B-GLMM are highlighted here, (a) item response order was related to response time, but not to item scores; and (b) item response time was related to item scores, but its effects were moderated by the cognitive level. Additionally, the findings from both data visualization and B-GLMM assisted instructors to see the alignment of student responses with purposefully designed item cognitive level and expected item response time.

We found that item response time was associated with item performance, and that this relationship was moderated by item cognitive level. This finding might not have found if the issue of outlying response times had not been resolved. As far as we know, this paper is the first attempt to address the issue of outlying response times that contained high level of measurement errors. We treated outlying response times as missing data and replaced them by multiple imputed values. Researchers can also explore other methods, such as robust estimator that can handle extreme non-normality.

Since item response order has not been studied in the literature, it is unknown whether the deviation of student response orders from the original item sequence can provide insightful information about students' test performance. Using data visualization approach, we observed that the more the order of the student responses deviated from the original item sequence, the more time was needed and lower scores were obtained. However, our results from the B-GLMM analyses did not find ROD to have a clear relationship with item scores, but it was positively related to item response time. Our results are based on one empirical data set, so there is a need for more studies to look into this issue. It is still uncertain why about 26% of the students did not follow the item sequence or why the rest did not choose to jump around when it was allowed. Based on our content experts' feedback, our findings may suggest that some students possibly did not have a good understanding on the course materials when they started to work on the quiz, so they had to take breaks to review the course materials or check their class notes.

Graphics is a great tool for classroom instructors to explore assessment data because it is easy to understand and interpret with no requirements on psychometric or statistical training. It provides a straightforward way of understanding data without any statistical assumptions. Visualization preserves the authenticity of the data and is not influenced by the statistical algorithms that might disguise or distort the original information in the data. The authentic characteristics of the data, such as low data variation (e.g., items #9 and #10 in **Figure 2**) or outliers (e.g., outlying response times in **Figure 4**), cannot prevent visualization from working properly as is often the case when a parametric statistical method is used. Data visualization is particularly useful when it is broken down by groups, such as low vs. high achieving, or English second language vs. native English Native speakers, which can help educators or researchers to target their teaching or research on specific groups of learners.

Graphics can also be very useful when exploring some unknown characteristics of online assessment data, such as item response order. In **Figure 4**, we found that a small proportion of students displayed dramatic deviation from the majority, who had relatively lower grades and used more time to answer quiz questions. The overall pattern of student performance can help to monitor whether the teaching objectives are achieved and identify areas that needs further classroom instruction or test revision. When under-performing or unusual learning behaviors of individuals are identified, intervention could take place in time to support student learning.

B-GLMM analysis is more flexible than the commonly used item response models, such as IRT models. De Boeck and Wilson (2004) indicated that both person and item property predictors could be included in the item response models to explain the person and item effects on item responses, which was known as doubly explanatory response models. In the data illustration, we demonstrated that B-GLMM allowed researchers to include both person (item response order, studying time and learning strategies) and item property (item cognitive level) predictors for understanding the item response time and item performance. Additionally, all the variables are allowed to be random or fixed effects. The addition of Bayesian approach to the generalized liner mixed model makes it less restricted to the requirement on sample size, eases the issues caused by model complexity, and helps to capture uncertain on the parameter estimates.

Using B-GLMM approach, the first model found that item response order and item cognitive level were related to item response time, which confirmed the observations obtained in **Figure 4** that showed that students in the high ROD group used much more time than those in the low ROD group. The second model confirmed our hypothesis that item cognitive level moderated the effect of item response time on item scores and also confirmed the findings from data visualization (**Figure 2**). The findings from B-GLMM suggest that it is appropriate to incorporate learning theory into data analysis if such theory is used for the assessment development.

Our study has two important practical implications to teaching and learning. First, the data visualization can be a practical tool for instructors to use for checking the alignment of student actual responses with the quiz design (i.e., item cognitive level, the expected item response time, and the expected understanding level). The observations based on data visualization can be confirmed by our B-GLMM analysis as we demonstrated in this study. The information obtained from item response time and response order adds additional dimensions to help instructors understand the cognitive effort students experienced with each item. If student responses on certain items deviate from the expectation (e.g., instructors thought this was an easy question, but it took students much more time), instructors can promptly start to review the item itself for flaws, or identify additional resources that can help students to improve their understanding. Additionally, we can share our findings on the item response order with the instructors and students, that is, students who had higher ROD values, took much more time to complete the quiz. The instructors could have a discussion with students about study habits and quiz-taking strategies, which may help students to improve their learning.

Although this research made a unique contribution to open online formative assessments, there are some limitations. First, most educators are not able to conduct the data analytical methods proposed in this study. It may be possible that we can make the data analysis automated though developing an online Shinny App or incorporating them with the learning management system. However, the interpretation of the results from B-GLMM analysis is not straightforward as data visualization, which requires psychometrician or statistician's extra help. There is a need to provide more practical methods for analyzing the data obtained from the open online formative assessments that are frequently used in classroom. Second, the use of one sample of undergraduate students at one university yields findings that are transferable only to a specific population. More studies with students from different disciplines and universities will help to provide a broad understanding. Finally, we only focused on a single quiz in this study. It will be even more informative, in future research, to model several quizzes simultaneously. Such approach can inform instructors and students the changes or patterns of student performance on the quiz items as well as their response behaviors over time.

The present study explored different strategies that are more suitable to study the characteristics of open online assessments as well as to handle relatively small sample size, like a regular class size (e.g., 100–300). The methods we proposed here can provide practical feedback to classroom instructors and facilitate their just-in-time teaching using item response time and response order in addition to item scores. We hope our study will motivate more research that explores diverse strategies for analyzing open online formative assessment data, which can inform teaching and learning.

DATA AVAILABILITY STATEMENT

The data presented in this paper are available upon the request. Please direct inquiries to yan.liu@ubc.ca.

AUTHOR CONTRIBUTIONS

YL contributed to the conception and design of the study, performed the statistical analysis, conducted part of the data visualization, and wrote the first draft of the manuscript. AB contributed to data visualization and the manuscript revision. AW contributed to the conception of the study and the manuscript revision. XZ contributed to the development of the new measure for quantifying item response order. CK contributed to the Bayesian data analysis and the manuscript revision. HB, MS, and JG contributed to the initiation of the research on open online assessments. HB developed the open online assessment and managed data collection. MS extracted assessment data from the Learning Catalytics platform and initiated the idea of using data visualization. JG contributed to the assessment design. All authors contributed to the article and approved the submitted version.

REFERENCES

Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Balch, W. R. (1989). Item order affects performance on multiple-choice exams. *Teach. Psychol.* 16, 75–77. doi: 10.1207/s15328023top1602_9

Baldwin, S. A., and Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychol. Methods* 18, 151–164. doi: 10.1037/a0030642

Biggs, J. B., and Tang, C. S. (2011). *Teaching for Quality Learning at University,* 4th Edn. McGraw-Hill, New York, NY: Society for Research into Higher Education & Open University Press.

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Computat. Graphic. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Brothen, T., and Wambach, C. (2001). Effective student use of computerized quizzes. *Teach. Psychol.* 28, 292–294. doi: 10.1207/S15328023TOP2804_10

Browne, W. J., and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* 1, 473–514. doi: 10.1214/06-BA117

Buchanan, T. (2000). The efficacy of a world-wide web mediated formative assessment. *J. Comput. Assist. Learn.* 16, 193–200. doi: 10.1046/j.1365-2729.2000.00132.x

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences,* 2nd Edn. L. Hillsdale, MI: Erlbaum Associates.

Davey, T., and Lee, Y. H. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE® revised general test. *ETS Res. Rep. Ser.* 2011, i—44. doi: 10.1002/j.2333-8504.2011.tb02262.x

De Boeck, P., and Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-BA117A

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Gelman, A., Carlin, J. B., Stern, H. S., Bunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis,* 3rd Edn. Boca Raton, FL: Taylor & Francis Group, LLC.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i02

Hambleton, R. K., and Traub, R. E. (1974). The effects of item order on test performance and stress. *J. Exp. Educ.* 43, 40–46. doi: 10.1080/00220973.1974.10806302

Ibabe, I., and Jauregizar, J. (2010). Online self-assessment with feedback and metacognitive knowledge. *Higher Educ.* 59, 243–258. doi: 10.1007/s10734-009-9245-6

Johnson, B. C., and Kiviniemi, M. T. (2009). The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teach. Psychol.* 36, 33–37. doi: 10.1080/00986280802528972

Kingston, N. M., and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Appl. Psychol. Meas.* 8, 147–154. doi: 10.1177/014662168400800202

Klein Entink, R. H., Fox, J. P., and van der Linden, W. J. (2008). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein, S., and Hamilton, L. (1999). *Large-Scale Testing: Current Practices and New Directions*. Pittsburgh, PA: RAND Corporation.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: an overview. *Theory Pract.* 41, 212–218. doi: 10.1207/s15430421tip4104_2

Laffitte, R. G. (1984). Effects of item order on achievement test scores and students' perception of test difficulty. *Teach. Psychol.* 11, 212–214. doi: 10.1177/009862838401100405

Lee, Y. H., and Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychol. Sci.* 53, 359–379.

Li, F., Cohen, A., and Shen, L. (2012). Investigating the effect of item position in computer-based tests. *J. Educ. Meas.* 49, 362–379. doi: 10.1111/j.1745-3984.2012.00181.x

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Appl. Psychol. Meas.* 34, 521–538. doi: 10.1177/0146621609355451

Meyers, J. L., Miller, G. E., and Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Appl. Meas. Educ.* 22, 38–60. doi: 10.1080/08957340802558342

Monk, J. J., and Stallings, W. M. (1970). Effects of item order on test scores. *J. Educ. Res.* 63, 463–465.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). *CODA: Convergence Diagnosis and Output Analysis for MCMC*. R News. Available online at: https://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf#page=7 (accessed December 12, 2020).

Rakes, G. C. (2008). Open book testing in online learning environments. *J. Interact. Online Learn.* 7, 1–9.

R-core (2018). *Package 'Parallel'*. 14. Available online at: https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf

Schnipke, D. L., and Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *J. Educ. Meas.* Hillsdale, MI 34, 213–232. doi: 10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D. L., and Scrams, D. J. (2002). "Exploring issues of examinee behavior: insights gained from response-time analyses," in *Computer-Based Testing: Building the Foundation for Future Assessments,* eds C. N. Mills, M. Potenza, J. J. Fremer, and W. Ward (Lawrence Erlbaum Associates), 237–266.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *Am. J. Polit. Sci.* 57, 748–761. doi: 10.1111/ajps.12001

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03

van de Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van de Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

Wang, T., and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Appl. Psychol. Meas.* 29, 323–339. doi: 10.1177/0146621605275984

Whitely, S. E., and Dawis, R. V. (1976). The influence of test context on item difficulty. *Educ. Psychol. Meas.* 36, 329–337. doi: 10.1177/001316447603600211

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wilson, M., De Boeck, P., and Carstensen, C. H. (2008). "Explanatory item response models: a brief introduction," in *Assessment of Competencies in Educational Contexts*, eds, J. Hartig, E. Klieme, and D. Leutner (Ashland, VA: Hogrefe & Huber Publishers), 91–120.

Wise, S. L., and DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *J. Educ. Meas.* 43, 19–38. doi: 10.1111/j.1745-3984.2006.00002.x

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Wu, L. (2009). *Mixed Effects Models for Complex Data*. New York, NY: Chapman and Hall/CRC.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a gibbs sampling approach. *J. Am. Stat. Assoc.* 86, 79–86. doi: 10.1080/01621459.1991.10475006

Zhao, Y., and Long, Q. (2017). Variable selection in the presence of missing data: imputation-based methods: variable selection in the presence of missing data. *Wiley Interdiscip. Rev.* 9:e1402. doi: 10.1002/wics.1402

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educ. Meas.* 10, 10–16. doi: 10.1111/j.1745-3992.1991.tb00198.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.