Received: 15 March 2022 | Accepted: 3 August 2022

DOI: 10.1111/bjet.13270

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Beyond item analysis: Connecting student behaviour and performance using e-assessment logs

Hatim Lahza[1,4] | **Tammy G. Smith**[2] | **Hassan Khosravi**[3]

[1]School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Queensland, Australia

[2]Office of Medical Education, The University of Queensland, St Lucia, Queensland, Australia

[3]Institute for Teaching and Learning Innovation, The University of Queensland, St Lucia, Queensland, Australia

[4]Department of Computer Science, College of Computers and Information Systems, Umm Al-Qura University, Mecca, Saudi Arabia

Correspondence

Hatim Lahza, School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia.
Email: h.lahza@uq.net.au

Abstract

Traditional item analyses such as classical test theory (CTT) use exam-taker responses to assessment items to approximate their difficulty and discrimination. The increased adoption by educational institutions of electronic assessment platforms (EAPs) provides new avenues for assessment analytics by capturing detailed logs of an exam-taker's journey through their exam. This paper explores how logs created by EAPs can be employed alongside exam-taker responses and CTT to gain deeper insights into exam items. In particular, we propose an approach for deriving features from exam logs for approximating item difficulty and discrimination based on exam-taker behaviour during an exam. Items for which difficulty and discrimination differ significantly between CTT analysis and our approach are flagged through outlier detection for independent academic review. We demonstrate our approach by analysing de-identified exam logs and responses to assessment items of 463 medical students enrolled in a first-year biomedical sciences course. The analysis shows that the number of times an exam-taker visits an item before selecting a final response is a strong indicator of an item's difficulty and discrimination. Scrutiny by the course instructor of the seven items identified as outliers suggests our log-based analysis can provide insights beyond what is captured by traditional item analyses.

KEYWORDS
assessment analytics, classical test theory, computer-based assessment, e-assessment, exam log, exam-taking behaviour, item analysis, learning analytics

Practitioner notes

What is already known about this topic

- Traditional item analysis is based on exam-taker responses to the items using mathematical and statistical models from classical test theory (CTT). The difficulty and discrimination indices thus calculated can be used to determine the effectiveness of each item and consequently the reliability of the entire exam.

What this paper adds

- Data extracted from exam logs can be used to identify exam-taker behaviours which complement classical test theory in approximating the difficulty and discrimination of an item and identifying items that may require instructor review.

Implications for practice and/or policy

- Identifying the behaviours of successful exam-takers may allow us to develop effective exam-taking strategies and personal recommendations for students.
- Analysing exam logs may also provide an additional tool for identifying struggling students and items in need of revision.

INTRODUCTION

Item analysis refers to a set of techniques that evaluate different characteristics of an assessment item, including its difficulty and discrimination, which can consequently be used to determine its effectiveness. One method of item analysis that has existed for several decades is classical test theory (CTT). Because CTT uses easily generated statistical analyses, it is commonly used by course instructors as a measure of the reliability of their exams. However, CTT suffers from two main drawbacks; it only considers exam-takers' performance to estimate item difficulty and discrimination abilities, and it is cohort oriented since it does not consider individual exam-taker characteristics.

While item analysis provides one way to interpret exam-takers' performance, it does not give us any insight into their exam-taking behaviour. Historically, most studies into predictors of success have either relied on the self-reported perceptions of the exam-takers (eg, when exploring behaviours and attitudes such as anxiety and motivation) or physical factors such as eraser marks and cross-outs when looking at answer changing patterns (Bauer et al., 2007; Couchman et al., 2016). However, other exam-taker behaviours such as the amount of time spent on each question, how many times they visit a question and the number of times they change an answer are more difficult or impossible to measure in paper-based exams.

In recent years, electronic assessment platforms (EAPs) that provide the ability for exams to be administered on- or off-line have become increasingly popular (Llamas-Nistal et al., 2013). One of the key advantages of EAPs is the ability to provide exam-takers with rapid and personalised feedback (Dennick et al., 2009). They also provide opportunities to deliver questions that would be difficult or impossible to deliver on paper, for example, questions incorporating multimedia. These new formats can be used to enhance the utility, reliability and validity of the assessment task (Dennick et al., 2009). An additional benefit of EAPs

is the ability to capture an exam-taker's journey through their exams. As EAPs have evolved, the data that can be extracted from each exam episode have become more sophisticated, allowing for scrutiny beyond the standard item analysis. One example is the snapshot feature within the ExamSoft EAP which records and timestamps every action and response made by an exam-taker throughout their exam. These snapshot files have a number of purposes, including exploring software bugs and investigating suspected academic misconduct.

This paper proposes a new approach inspired by learning analytics (LA) methods to investigate whether logs generated by EAPs can be used to approximate and/or complement the difficulty and discrimination indices of assessment items obtained by CTT. Our approach entails learning analytics and educational data mining methods, particularly feature engineering, statistics, visualisation and outlier analysis. In particular, it aims to contribute to the literature by (1) presenting techniques that approximate an item's difficulty and discrimination based on exam-taker behaviour, (2) presenting techniques that complement CTT by identifying items that may require instructor review and (3) evaluating our proposed approaches using real data obtained from an EAP.

BACKGROUND AND RELATED WORK

Item analysis

Classical test theory

The two primary measures of item analysis are the difficulty and discrimination indices (De Champlain, 2010). These indices can assist examiners in determining whether items are functioning as intended. The item difficulty index is referred to as the *p*-value and is generally calculated as the proportion of exam-takers who answered the question correctly, with a possible range of 0.00–1.00 (the closer to 1.00, the easier the item). It is important to remember that the *p*-value is not only representative of the difficulty of the content but also the ability level of the cohort being tested (De Champlain, 2010). It is also useful to assess whether exam items discriminate between exam-takers of differing abilities within the same cohort. We would generally expect a higher proportion of more capable exam-takers to correctly answer a given item than those struggling (De Champlain, 2010). The discrimination index may be calculated by subtracting the *p*-value of the lower 27% of the exam-takers from the *p*-value of the upper 27%. The possible range for the discrimination index is −1.00 to 1.00, where a score close to 0 indicates that the higher and the lower performing exam-takers scored similarly on the item. In general, there are no standard cut-off values to determine discrimination ability (Chiavaroli & Familari, 2011). While it is usually desirable for an item to have a positive discrimination index, it is also acceptable to have some items that target threshold concepts (Cousin, 2006) that have a discrimination index of close to 0. Therefore, interpretation of the quality of an item based on its discrimination index still requires academic judgement rather than being an automated process.

Advanced methods

One limitation of CTT is that it is cohort oriented, meaning that the estimation of item indices depends on the abilities of the exam-takers of a particular sample. This limitation is addressed by complex modelling of item responses to estimate latent traits that might represent an exam-taker's ability. Traditionally, item response theory (IRT) models the correctness of responses considering their association with exam-taker ability θ and sometimes a number

of item characteristics (Livingston, 2006). In other words, the probability of getting the correct answer is a function of θ that can be represented as a logistic curve (S-shaped) where exam-takers with lower abilities will have probabilities close to zero, and those with higher abilities will have the highest probabilities. In addition, when item characteristics are assumed to affect the estimation, difficulty β and discrimination α parameters are used. The parameter β represents θ at the centre of the curve, while α represents how the changes in the abilities are reflected in the changes of the estimation values (slope at the centre of the curve) (Wang & Bao, 2010). There have been many extensions over the classical item response theory in the context of adaptive educational systems that automatically recommend learning resources and instructions to students (Abdi et al., 2019, 2020; Abdi, Khosravi, & Sadiq, 2021; Abdi, Khosravi, Sadiq, & Darvishi, 2021b; Lee, 2019; Wauters et al., 2010). Despite the superior performance of these models, they have some limitations that limit their adoption over CTT for item analysis. Firstly, they are mathematically more complex which makes them harder to understand. Secondly, they often require multiple and large sample sizes for calibration, which may not always be available. In addition, they still only use exam-taker responses to items as input and do not have the capacity to easily consider exam-takers' behaviour, as we will in this paper, in analysing items.

Log files in electronic assessment platforms

To our knowledge, the idea of using log data from an EAP to analyse exam items was first introduced by Neel's 1999 work, presented at the Annual Meeting of AERA (cited in Jung Kim, 2001). To date, exam logs have mostly been used for measuring and modelling exam-takers' accuracy, speed, revisits and effort (Bezirhan et al., 2021; Klein Entink et al., 2008; Sharma et al., 2020; Wise, 2015; Wise & Gao, 2017); analysing answering and revising behaviour during exams (Costagliola et al., 2008; Pagni et al., 2017); examining and enhancing metacognitive regulation of strategy use and cognitive processing (Dodonova & Dodonov, 2012; Goldhammer et al., 2014; Papamitsiou & Economides, 2015; Thillmann et al., 2013); classifying exam-takers towards testing services personalisation (Papamitsiou & Economides, 2017); validating the interpretations of test score (Engelhardt & Goldhammer, 2019; Kane & Mislevy, 2017; Kong et al., 2007; Padilla & Benítez, 2014; Toton & Maynes, 2019; van der Linden & Guo, 2008); understanding exam-takers' performance (Greiff et al., 2016; Kupiainen et al., 2014; Papamitsiou et al., 2014, 2018; Papamitsiou & Economides, 2013, 2014); enhancing item selection in adaptive testing environment (van der Linden, 2008); analysing exam items (Costagliola et al., 2008; Jung Kim, 2001); detecting cheating (Cleophas et al., 2021; Costagliola et al., 2008); and identifying test-taking strategies (Costagliola et al., 2008). Nonetheless, most of the previous work focused on time-based behaviours and the interpretation of exam-taker results; few of them examined the potential of using exam-taker behaviours to validate or enrich the interpretation of the quality of exam items. This paper focuses on investigating whether logs generated by an EAP can be used to approximate the difficulty and discrimination level of assessment items and if so whether this information can be used towards identifying items that may require instructor review.

Linking learning analytics and assessment

Lang et al. argue that ‘by design—or else by accident—the use of a learning analytics tool is always aligned with assessment regimes, which are in turn grounded in epistemological assumptions and pedagogical practices’ (Lang et al., 2017, p.13). Hence, it comes as

no surprise that researchers have discussed the benefits of the synergy between the two fields on several levels of the educational system (eg, C. Ellis, 2013; Ifenthaler et al., 2018; Jordan, 2013). On the learner level, learners can monitor their progress in real time with evidence and receive a personalised learning experience. On the teaching staff level, instructors can attain information from assessment platforms to improve course planning, procedure and policy. They can also gain in-depth and fine-grained insights into learning processes and behavioural patterns and provide support and personalised interventions, especially in large-scale assessments. On the institutional level, educational institutions can perform annual course and module evaluations in comparison with other institutions or schools. In addition, learning analytics can improve assessment practice and learning designs (Barana et al., 2019) by testing open hypotheses (Gašević et al., 2022), move the focus from the product (summative) to the process led to that product (Palmiero & Cecconi, 2019) and advance fairness and bias concerns in education (Gašević et al., 2022). However, only few research has investigated how advancement in learning analytics can benefit assessment research and practice. Most recently, genuine and solid research endeavours have been established to link (Ifenthaler & Greiff, 2021) and explore how to strengthen the links between learning analytics and assessment (Gašević et al., 2022). A recent special issue of Computers and Human Behaviour introduced a number of peer-reviewed articles which strengthened the links between learning analytics and assessment (Gašević et al., 2022). The editors of that special issue grouped those articles into three broad categories: (1) analytics for assessment (ie, learning analytic approaches to support assessment); (2) analytics of assessment (ie, investigating assessment practices and properties of assessments); and (3) validity of measurement (ie, concerning the validity in measurement in learning analytics). However, the editors also stated that these three categories were not intended to cover every possible link between those two fields. Hence, the presented study might not clearly fit into only one of these categories.

METHOD

Research questions

Our exploration is guided by the following problem research questions. Given a log file L that captures the timestamped steps an exam-taker takes while completing an exam, exam grades G and CTT-based difficulty index (Diff-I) and discrimination index (Disc-I) of the exam items, we focus on answering the following questions:

1. Are exam-taker behaviours in engaging with an assessment item indicative of its difficulty level?
2. Are exam-taker behaviours in engaging with an assessment item indicative of its discrimination level?
3. Can exam-takers behaviours complement CTT item analysis in identifying items that may require instructor review?

Approach

In this section, we present our approach to answering the research questions provided in Section 3.1. To answer RQ1 and RQ2, we extract features representative of exam-taker behaviours, then check their fit and rank them using our available data (ie, CTT item analysis). To answer RQ3, we use outlier analysis using a regression technique. Algorithm 1

provides a high-level pseudocode for our approach. Steps one and two discuss feature selection, while steps three to five are analytically driven and are used to answer and report the findings of our proposed RQs and are also discussed in the Results section. An actual implementation in R alongside a sample data set representing our input data is released on GitHub at https://github.com/hlahza/BJET_beyond-item-analysis.

Algorithm 1: Identifying exam items that may require instructor review

Input : L, D, C, G [L: **log**, D: **Diff-I**, C: **Disc-I**, G: **grades**];
Output : Σ [**A set of items needing review**];
1 $examTakerItemFeature \leftarrow computeExamItemFeature(L)$ [**Compute each exam-taker/question feature using the definitions in Table 1 as described in Section 3.2.1**];
2 $diffItemLevelFeature, discItemLevelFeature \leftarrow computeItemLevelFeature(examTakerItemFeature)$ [**Compute item-level feature according to the definitions in Section 3.2.2**];
3 $diffRanking, discRanking \leftarrow rankFeatures(diffItemLevelFeature, discItemLevelFeature, D, G)$ [**Rank the Diff-F and Disc-F features based on the correlation (r) as described in Section 3.2.3**];
4 $\Sigma \leftarrow detectOutliers(diffRanking, discRanking, D, C)$ [**Detect outliers using regression models between Diff-I/Disc-I and Diff-F/Disc-F and standardised residuals as described in Section 3.2.4**];
5 **return** Σ;

Exam-taker/item pair feature descriptions and definitions

An important initial step in this investigation is transforming the raw data from an exam event log *L* into meaningful features that represent exam-taker behaviours during an examination. Hence, the first step in Algorithm 1 is to extract exam-taker/item pair features. In this study, the selected features are aimed at capturing exam-takers' behaviours in both *answering* and *reviewing* assessment items, although there is a clear overlap between these two activities. Answer changing, item visiting and time spent on a question were the initial behaviours identified by the authors, as they frequently appeared in previous research. This list was expanded through several brainstorming sessions by considering different approaches exam-takers may adopt while navigating through an exam; for instance, skipping difficult questions and returning to them at the end of the exam or reviewing confusing exam questions before submitting the exam. The final list of 11 exam-taker/item pair features summarised in Table 1 was categorised into two groups: seven *frequency-based* features and four *time-based* features. Frequency-based features report the number of times an event has occurred (minimum = 0), whereas the time-based feature reports the amount of time (in seconds) that was spent on an event.

To compute these features, we track exam-taker interactions with an item into different blocks, where each block contains a set of actions taken for that item. We refer to each of these blocks here as visits (v), which can also be thought of as the time spent navigating to an item, performing some action (eg, answering or marking with a flag) and then navigating to another item. To illustrate, we use an example from our available data used to evaluate the approach in Section 3.3. Figure 1 shows the specific example of one exam-taker's interaction with item 33 (which, for this exam-taker, was preceded by item 50). In this example, we identified five blocks (labelled $v1$ to $v5$).

Figure 2 below provides a graphical depiction of how the features are computed based on these blocks. The top part of the figure demonstrates how frequency-based features are computed and the bottom part of the figure demonstrates how time-based features are computed. A more detailed written description of how each of the frequency- and time-based features is computed, with reference to Figures 1 and 2, is provided in the supplementary on-line material (Box S1).

TABLE 1 Brief description of time- and frequency-based exam-taker/item pair features

Type	Abbreviation	Feature	Description
Frequency-based	IA	Item actions	Total number of actions
	AC	Answer changing	NT an exam-taker changed their response
	IV	Item visiting	NT an exam-taker visited an item
	IV.BIS	Item visiting before initial selection	NT an exam-taker visited an item before selecting initial response
	IV.FIS	Item visiting following initial selection	NT an exam-taker visited an item after selecting initial response
	IV.BFS	Item visiting before final selection	NT an exam-taker visited an item before selecting final response
	IV.FFS	Item visiting following final selection	NT an exam-taker visited an item after selecting final response
Time-based	AT	Answering time	TS before final response selection
	RT.FIS	Review time following initial selection	TS after selection of the initial response
	RT.BFS	Review time before final selection	TS between selection of initial and final response
	RT.FFS	Review time following final selection	TS after selection of the final response

Abbreviations: NT, number of times; TS, time spent.

	Seq	Item#	Snapshot#	Item Type	Timestamp	Trigger	Response
*V*1	114	50	1	Choice	8:39:34 am	Navigation	
	115	33	1	Choice	8:39:38 am	Navigation	
					.....		
*V*2	239	50	2	Choice	9:10:17 am	Navigation	
	240	33	2	Choice	9:10:20 am	Answered	Choice(s): A
	241	33	3	Choice	9:10:24 am	Navigation	
					.....		
V3	448	50	3	Choice	9:49:53 am	Navigation	
	449	33	4	Choice	9:49:54 am	Answered	Choice(s): B
	450	33	5	Choice	9:49:55 am	Answered	Choice(s): C
	451	33	6	Choice	9:49:58 am	Answered	Choice(s): A
	452	33	7	Choice	9:49:59 am	Navigation	Choice(s): A
					.....		
*V*4	578	50	3	Choice	10:02:50 am	Navigation	Choice(s): C
	579	33	8	Choice	10:02:51 am	10 Minute Save	Choice(s): A
	580	33	9	Choice	10:02:52 am	Navigation	Choice(s): A
					.....		
*V*5	620	50	4	Choice	10:09:51 am	Navigation	Choice(s): C
	621	33	10	Choice	10:09:54 am	Navigation	Choice(s): A
					.....		

FIGURE 1 A screenshot of one of the exam-taker snapshots for one item.

Item-level feature descriptions and definitions

Once the exam-taker/item pair features are extracted for the entire cohort, step 2 of Algorithm 1, where two sets of item-level features (Diff-F and Disc-F) based on Diff-I and Disc-I are generated, should be implementd. The patterns (Diff/Disc-F) can be read as difficulty/discrimination-based features where any particular feature in Table 1 can substitute the letter F. Diff-F features are generated by calculating the mean of the entire cohort for each of the frequency- and time-based features described in Table 1. For example, Diff-AC

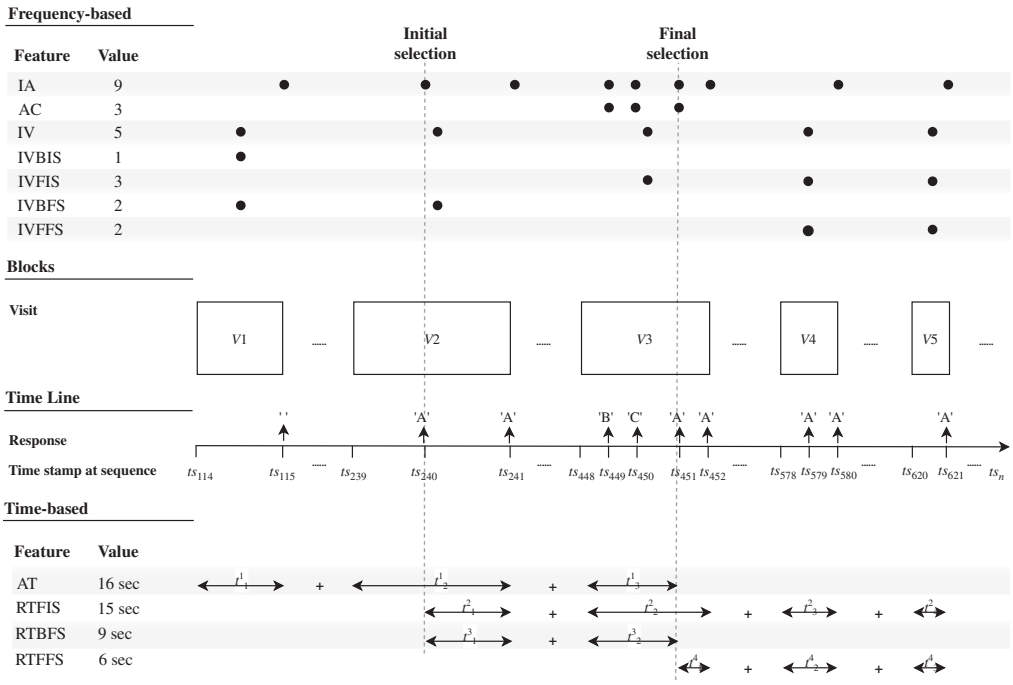


FIGURE 2 Visual demonstration of the features. ts_i, V and t stand for timestamp at sequence i, visit and time respectively.

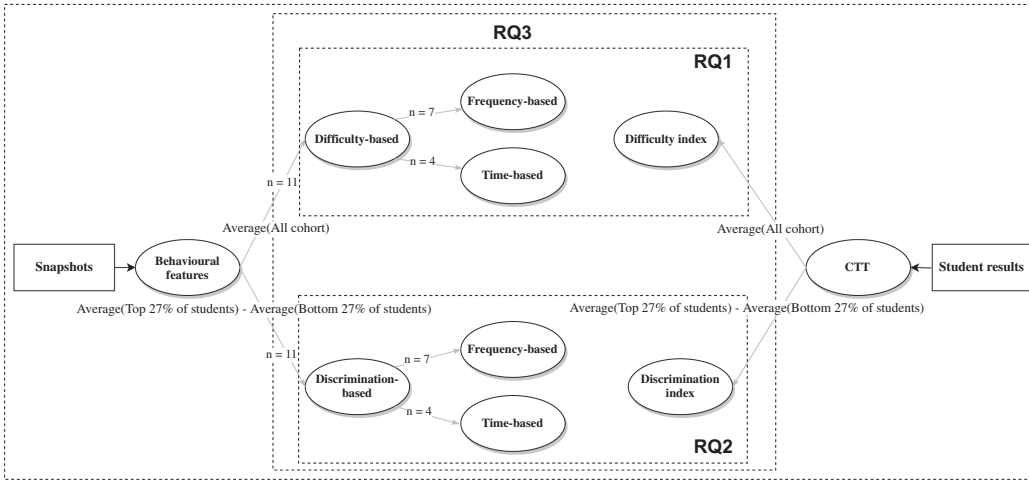


FIGURE 3 Overview of the item-level features in comparison to CTT item indices.

is the average number of changes made for each item by the entire exam-taker cohort. Disc-F features are based on the difference between the upper and lower performing 27% of exam-takers for each of the frequency- and time-based features described in Table 1. For example, the Disc-AC is calculated by subtracting the average number of changes per item of the lower 27% of exam-takers from the average of the upper 27% of exam-takers. In total, 22 item-level features should be generated (11 difficulty-based and 11 discrimination-based). This process and the relationship to our research questions are illustrated in Figure 3.

Ranking features

Once the 22 item-level features are generated, we look for correlations and differences between each feature and the corresponding CTT index. The correlation and regression analysis are used to find the features that best fit traditional item analysis indices. In step 3 of Algorithm 1, the resultant correlations are used to rank the item-level features on Diff-F and Disc-F (ie, the higher the correlation coefficient, the higher the rank). These rankings are used for answering RQ1 and RQ2 as presented in Sections 4.1 and 4.2 respectively. In addition, to gain a deeper understanding of the relationship, the Wilcoxon–Mann–Whitney U test of independence is used to examine the differences between the medians of the behavioural features per CTT item analysis categories.

Outlier detection

In Section 3.2.2, we have identified 22 item-level features. Here, we assume that the behaviour that generates a pair of features Diff-F and Disc-F (eg, Diff-IV and Disc-IV) with the highest correlation with Diff-I and Disc-I holds the most information about exam items' difficulty and discrimination. Consequently, to implement step 4 from Algorithm 1, we first calculate the average ranking of the ranks generated in Section 3.2.3 to determine the best pair of features that fits Diff-I and Disc-I. Second, inspired by the common analysis conducted in CTT for investigating the relationship between Diff-I and Disc-I (Aiken, 1979; Hingorjo & Jaleel, 2012; Karelia et al., 2013; Sim & Rasiah, 2006), we identify outlier items by fitting polynomial regression models on Diff-I and Disc-I, Diff-I and Diff-F and Disc-I and Disc-F. Third, an item is deemed to be an outlier if it has a standardised residual $> 2.5 \vee < -2.5$from the fitted regression line. More details on interpreting the models and a review by a medical education specialist are provided in the Result Section 4.3.

Evaluation

To test our approach, we used de-identified assessment data from 463 medical students enrolled in a first-year biomedical sciences course. The exam analysed in this article initially consisted of 90 multiple choice questions (MCQs), each with five options, from which exam-takers were asked to select the single best answer. Of these initial 90 items, one was removed from our analysis as it was identified by course personnel as having no truly correct answer. Exam-takers sat the exam electronically on the Examplify app (part of the ExamSoft EAP) on their own or institutional devices. Questions were presented in random order to each exam-taker. Three data sets generated by the ExamSoft EAP were analysed for this study: snapshot files, individual exam-taker results and item difficulty and discrimination indices (item analysis).[1]

While individual exam-taker snapshots are readily accessible to examiners, the system does not support the bulk download of these files for large numbers of exam-taker. For the purposes of this study, we were provided by the ExamSoft support team with a comma-separated value (CSV) file containing the following whole-of-cohort data:

- Item number: the question number as it appears on the exam.
- Student sequence: the order in which the questions were delivered to an exam-taker.
- Snapshot number: the number of actions received by the question.
- Item type: 'Choice' for an exam consisting solely of MCQs.

- Timestamp: the current time of the local machine when the exam-taker took an action, for example, navigating from one question to another.
- Trigger: the type of action performed by the exam-taker or the platform.
- Response: the exam-taker's answer to a question.

For our study, a difficult question (n = 0) was defined as an item with a Diff-I of <0.25 whereas an item with a Diff-I of >0.75 was considered an easy question (n = 48, 53.9%). Items with a Diff-I of between 0.25 and 0.75 were classified as being moderately difficult (n = 41, 46.1%). An item with a Disc-I of ≥0.2 was considered to be a good discriminator (ie, have reasonable ability in discriminating between exam-takers of differing ability) (n = 44, 49.4%). Items with a Disc-I of <0.2 were classified as poor discriminators (n = 45, 50.6%).

RESULTS

Response to RQ1: Difficulty-based features versus difficulty index

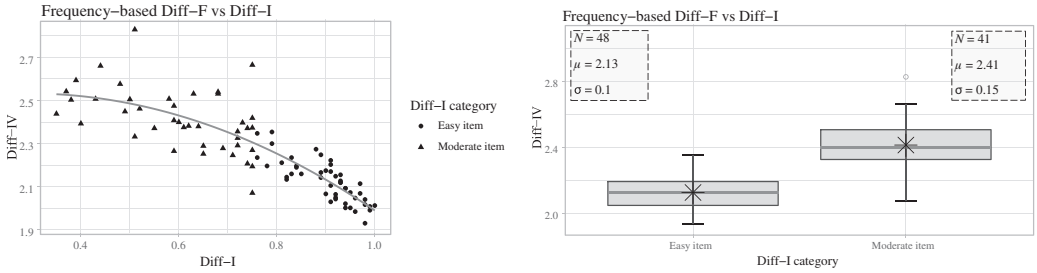
This section reports our findings on RQ1 and whether exam-taker behaviours in engaging with an assessment item are indicative of its difficulty level. Table 2 lists and places in rank order the Spearman correlation between the difficulty-based item-level features and the Diff-I. It is important to recall that the *higher* the Diff-I, the *easier* the item; therefore, when a feature is *negatively* correlated with Diff-I, it indicates a positive correlation with item difficulty. For example, the more difficult the item, the greater the number of visits and the more time spent on reviewing before the final selection is made.

Figure 4a illustrates the item-level relationship between Diff-I and Diff-IV, as the selected feature with the highest correlation with Diff-I. A regression model has been added to the graph (polynomial regression: $F[2, 86] = 133.1$, $p < 0.01$, with R^2 of 0.75) to demonstrate the trend of the data. In Figure 4b, we use a box and whisker plot to further illustrate this relationship by grouping items according to their Diff-I category. The Wilcoxon rank-sum test indicates that the difference between these two groups is statistically significant (Z = 82, p < 0.01, two-tailed test), supporting the relationship between item difficulty and item visits.

TABLE 2 Correlation between the difficulty-based item-level features (diff-F) and difficulty index (diff-I)

Rank	Feature	Type	Spearman
1	Diff-IV	Frequency-based	−0.90**
2	Diff-IV.BFS	Frequency-based	−0.89**
3	Diff-IV.FIS	Frequency-based	−0.89**
4	Diff-IV.BIS	Frequency-based	−0.85**
5	Diff-AC	Frequency-based	−0.85**
6	Diff-IA	Frequency-based	−0.82**
7	Diff-RT.BFS	Time-based	−0.80**
8	Diff-IV.FFS	Frequency-based	−0.77**
9	Diff-RT.FIS	Time-based	−0.70**
10	Diff-AT	Time-based	−0.59**
11	Diff-RT.FFS	Time-based	−0.54**

**Indicates correlations that are significant at 0.001 level (two-tailed). Note, in the feature names, BFS stands for before final selection, FIS stands for following initial selection, BIS stands for before initial selection and FFS stands for following final selection.



(a) The relationship between Diff-IV and Diff-I. (b) Mean Diff-IV per difficulty category.

FIGURE 4 Item visiting behaviour in relation to difficulty index.

TABLE 3 Correlation between the discrimination-based item-level features (disc-F) and discrimination index (disc-I)

Rank	Feature	Type	Spearman
1	Disc-IV.BFS	Frequency-based	−0.53**
2	Disc-IV.BIS	Frequency-based	−0.50**
3	Disc-IV.FFS	Frequency-based	0.45**
4	Disc-RT.FFS	Time-based	0.27*
5	Disc-AC	Frequency-based	−0.25*
6	Disc-AT	Time-based	−0.23*
7	Disc-IV.FIS	Frequency-based	0.22*
8	Disc-IA	Frequency-based	−0.16
9	Disc-RT.BIS	Time-based	0.12
10	Disc-IV	Frequency-based	−0.04
11	Disc-RT.BFS	Time-based	−0.02

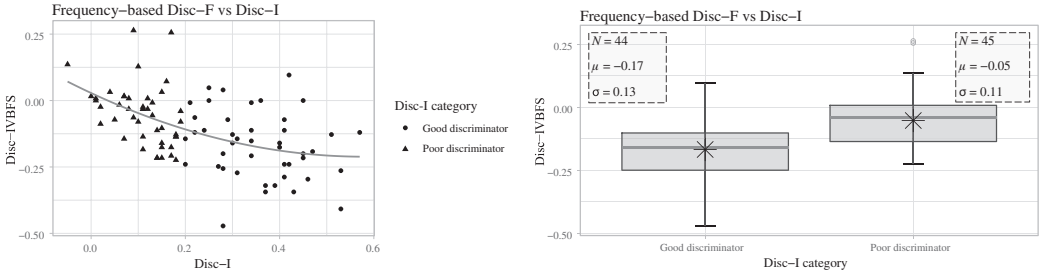
*Indicates that correlation is significant at the 0.05 level (2-tailed). Note, in the feature names, BFS stands for before final selection, FIS stands for following initial selection, BIS stands for before initial selection and FFS stands for following final selection.

**Indicates that correlations are significant at 0.001 level (2-tailed).

Response to RQ2: Discrimination-based features versus discrimination index

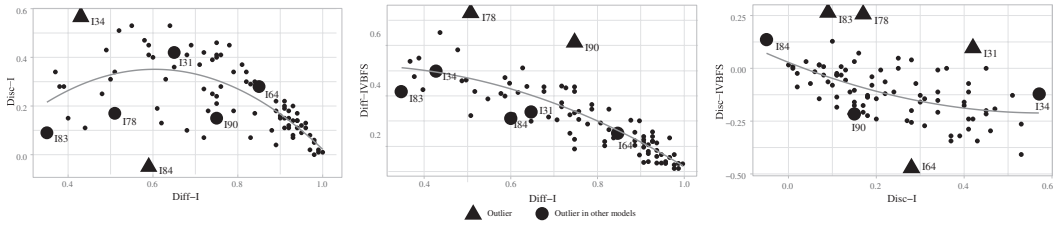
This section reports our findings on RQ2 and whether exam-taker behaviours in engaging with an assessment item are indicative of its discrimination level. Table 3 lists and places in rank order the Spearman correlation between the discrimination-based item features and the item discrimination index.

Figure 5a illustrates the item-level relationship between Disc-I and Disc-IV.BFS, as the selected feature with the highest correlation with Disc-I. A regression model has been added to the graph (polynomial regression: $F[2, 86] = 16.22$, $p < 0.01$, with R^2 of 0.27) to demonstrate the trend of the data. In Figure 5b, we use a box and whisker plot to further illustrate this relationship by grouping items according to their ability to discriminate (good and poor). The Wilcoxon rank-sum test shows the difference between these two groups is statistically significant ($Z = 508.5$, $p < 0.01$, two-tailed test) and supports the hypothesis that there is a relationship between the discrimination index of an item and the number of visits made before making a final selection.



(a) The relationship between Disc-IV.BFS and Disc-I. (b) Mean Disc-IV.BFS per discrimination category.

FIGURE 5 Item visiting before final selection behaviour in relation to CTT discrimination index.



(a) Diff-I vs Disc-I. (b) Diff-I vs Diff-IV.BFS. (c) Disc-I vs Disc-IV.BFS.

FIGURE 6 Quadratic regression models for Disc-I versus Diff-I, Diff-I versus Diff-IV.BFS and Disc-I versus Disc-IV.BFS.

Response to RQ3: Identification of outliers

This section reports our findings while investigating RQ3 to determine whether exam-taker behaviours can complement CTT item analysis in identifying items that may require instructor review. At the end of the section, in Table 5, we provide a summary of the insight gained from the analysis and the educator of the course.

Figure 6a illustrates the relationship between the Diff-I and Disc-I of the items in our study. The solid line demonstrates the best polynomial regression model fit to these data, which suggests that the relationship between Diff-I and Disc-I of assessment items is not linear but more dome-shaped. This finding is supported by past studies (eg, Sim & Rasiah, 2006) and the following theoretical justification. Questions with Diff-I ≈ 0 (ie, no or very few exam-takers answer the question correctly) or with Diff-I ≈ 1 (ie, all or most exam-takers answering the question correctly) both have a Disc-I ≈ 0 as they have very little power in discriminating between upper and lower performing exam-takers. Questions with $0.3 \leq$ Diff-I ≤ 0.7 (ie, where there are a significant number of correct and incorrect answers) can have a larger Disc-I as they have the capacity to discriminate between upper and lower exam-takers. A potential benefit of analysing the relationship between Diff-I and Disc-I is to identify outlier items which may require more detailed scrutiny by the examiners. As commonly used in regression analysis, we used the geometric distance between the fitted line and observed values to identify outliers (Wiggins, 2000). In our analysis, items with a standardised residual > 2.5 (Items 34 and 84) were considered outliers. These points are represented by triangles and labelled in Figure 5a. The larger circles indicate items that were flagged as outliers using the alternative feature-based approaches outlined below.

Two alternative approaches for identifying outlier items are to examine the relationship between the difficulty-based item-level features (Diff-F) and the difficulty index (Diff-I) (based on RQ1) and between the discrimination-based item-level features (Disc-F) and the

discrimination index (Disc-I) (based on RQ2). In the following section, we investigate how identifying outliers from these two approaches can complement the outlier detection analysis based on CTT. The feature we selected for this analysis was the one with the best average rank across both Diff-F and Disc-F, which is IV.BFS as indicated in Table 4.

Figure 6b illustrates the relationship between Diff-I and Diff-IV.BFS of items in our study. The solid line demonstrates the best polynomial regression model. Items with a standardised residual $>$ 2.5 (items 78 and 90) were considered outliers. These points are represented by triangles and labelled with the item number. The larger circles indicate items that were flagged as outliers using the two alternative approaches.

Similarly, Figure 6c illustrates the relationship between Disc-I and Disc-IV.BFS of items in our study. The solid line demonstrates the best polynomial regression model. Items with a standardised residual $>$ 2.5 (items 83, 78, 64 and 31) were considered outliers. These points are represented by triangles and labelled with the item number. The larger circles indicate items that were flagged as outliers using the two alternative approaches.

Based on the three regression models described above, seven items (31, 34, 64, 78, 83, 84 and 90) were identified as outliers and subject to review by a medical education specialist (see Table 5).

DISCUSSION

Exam-taker behaviours such as item skipping, answer changing, item-revisiting and time spent on an item, may reflect multiple latent factors (eg, cognitive ability, anxiety, engagement, confidence, familiarity of the content and item-design characteristics). An exam-taker's knowledge about the subject matter could be reflected in the time they take to answer the question or the number of times they visit an item. A confident exam-taker will choose the correct response and move on to the next question whereas a less confident exam-taker may either spend longer on the question or skip it altogether (Stenlund et al., 2018); behaviours also known to be associated with during-test test-taking strategies (Ellis & Ryan, 2003; Hong et al., 2006; Stenlund et al., 2017). When exam-taker behaviours are analysed collectively,

TABLE 4 Average rank of attributes across both difficulty-based (diff-F) and discrimination-based (disc-F) item-level features

	Diff-F-based version		Disc-F-based version		
Feature name	Spearman	Rank	Spearman	Rank	Average rank
IV.BFS	−0.89	2	−0.53	1	1.5
IV.BIS	−0.85	4	−0.50	2	3
AC	−0.85	4	−0.25	5	4.5
IV.FIS	−0.89	2	0.22	7	4.5
IV.FFS	−0.77	8	0.45	3	5.5
IV	−0.90	1	−0.04	10	5.5
IA	−0.81	6	−0.16	8	7
RT.FFS	−054	11	0.27	4	7.5
AT	−0.59	10	0.23	6	8
RT.FIS	−0.70	9	0.12	9	9
RT.BFS	−0.80	7	−0.02	11	9

Note: In the feature names, BFS stands for before final selection, FIS stands for following initial selection, BIS stands for before initial selection and FFS stands for following final selection.

TABLE 5 Outlier items with a summary of insights gained

Item #	Outlier			Summary of findings	Educational review of item
	CTT	Diff	Disc		
31	No	No	Yes	The upper 27% of exam-takers made a higher than expected number of visits and the lower 27% of exam-takers made a lower than average number of visits	Exam-takers would either know the answer to this recall question or not. Poorer exam-takers would gain nothing by returning to it even if they were unsure of the answer and may have simply 'cut their losses' and decided not to waste time on revisiting the question
34	Yes	No	No	This item had a higher Disc-I than predicted. It also recorded the second highest number of changes of all items	All of the options in this item were potentially correct, with exam-takers being required to select the most important option. A student who was less confident of the correct answer would understandably jump between responses
64	No	No	Yes	The upper 27% of exam-takers made very few visits to the item while the lower performing exam-takers made many more visits than expected	It appears that the upper 27% of exam-takers were confident of their response and did not require a second visit. Of the lower 27% of exam-takers, almost all who chose incorrectly eventually chose the same distractor, which was plausible from the context of the question but wrong
78	No	Yes	Yes	The average number of visits was significantly higher than expected and the upper 27% of exam-takers visited more often than the lower 27% of exam-takers. This item also had the third highest number of answer changes of all items in the exam with the upper exam-takers changing their answers almost twice as often as the lower exam-takers	This question required careful interpretation and clinical reasoning. It is possible that the poorer exam-takers misinterpreted the question and were confident in their (incorrect) response and so did not revisit, whereas the upper exam-takers revisited to ensure their interpretation was correct and subsequently changed their answers
83	No	No	Yes	The upper 27% of exam-takers made a higher than average number of visits and the lower 27% of exam-takers made a lower than average number of visits	This was arguably a subjective question based on very specific information provided in a lecture. Exam-takers selected from the full range of options so had a high level of uncertainty. Only 42% of the upper 27% of exam-takers answered the question correctly. This item should be revised
84	Yes	No	No	This moderately difficult item had an unpredictably very poor Disc-I More of the lower 27% of exam-takers answered this item correctly than the higher performing exam-takers. The upper 27% also made more visits to the item than the lower performing exam-takers and changed their answer at a much higher rate	On face value, this was a relatively easy question; however, we can see that the higher performing exam-takers found it more difficult than their lower performing peers. A possible explanation was that the upper exam-takers were overthinking what was a relatively straightforward question (looking for a trick perhaps) and were reluctant to choose 'normal age-related changes' over a pathological condition

TABLE 5 (Continued)

Item #	Outlier			Summary of findings	Educational review of item
	CTT	Diff	Disc		
90	No	Yes	No	On average, this item received a higher number of visits before final selection than predicted yet the difference between upper and lower exam-takers was as predicted This item had the ninth highest number of changes of all items, although unlike Item 78, the lower exam-takers changed their answers at a greater frequency than the upper exam-takers	From an educational perspective, this was a two-stage clinical question where exam-takers had to first identify a condition based on a scenario, and then choose the most appropriate drug to treat it. It is apparent that regardless of their level of achievement, exam-takers overestimated this item's difficulty based on the greater effort expended

they can be used as powerful cues of those latent factors which in turn can supply information beyond that of the traditional item analysis.

This paper adds to the literature by demonstrating that the data obtained from computer-based assessment logs can be used to complement and extend traditional CTT item analysis. We have identified several features that strongly correlate with difficulty and discrimination indices and have additionally detected outlier questions not picked up by traditional item analysis. We have also been able to refer back to the original exam questions to seek to understand why exam-taker behaviours may have varied from that predicted for certain items.

All 11 of the behavioural features that we selected had a significant negative correlation with Diff-I; that is the more difficult the question, the greater the incidence of the frequency-based behaviours and the longer the times for the time-based behaviours. It makes perfect sense that a more difficult question would require more answering effort on behalf of the exam-takers.

It is interesting to further analyse two behaviours that, at face value, look quite similar (IV and IV.BFS). Here, we see markedly different results when exploring relationships between these behaviours and Diff-I and Disc-I. Both the total number of times an exam-taker visited an item (IV) and the number of visits they made before selecting their final response (IV.BFS) strongly correlated with the difficulty of the question. However, while there was a significant negative correlation between IV.BFS and Disc-I (indicating that the lower performing exam-takers required significantly more visits before selecting their final response than the upper exam-takers for questions of good discrimination), there was no correlation between total item visits and item discrimination (indicating that upper and lower exam-takers visited at similar rates regardless of the Disc-I of the item).

We believe this relates to purposeful visiting; visits made following a final selection may simply indicate that an exam-taker is filling in time while waiting for the exam to finish rather than being an indication of uncertainty. This is supported by the positive correlation between Disc-IV.FFS and Disc-I, indicating that the upper exam-takers made more visits after selecting their final responses that the lower exam-takers for items of good discrimination. In other words, IV.BFS is more likely to represent confidence in a response; the fewer visits, the more confident the exam-taker is of their response.

Interestingly, we can see from Tables 2 and 3 that frequency-based features have a stronger correlation with the CTT indices than time-based features. We believe this is because the time a student spends on an item is influenced not only by the difficulty of the

item but also by the length and/or complexity of the question, whereas this is less likely to affect the frequency of visits to the item. Time-based features are also more likely to be influenced by how an exam-taker manages their time during a test; for example, an exam-taker who finishes their exam with time to spare may choose to upload their answer file immediately on completion, whereas another student may keep reviewing items until the end of the exam period.

By identifying outliers based on exam-taker behaviour, we have extended the knowledge that can be obtained by CTT alone. For example, how does an item with a low Diff-I and a low visiting frequency differ from an item with a low Diff-I but a high visiting frequency? Both items may be poor or incorrect, but for various reasons. Exam-takers seemed confident of the former item (not revisiting it), so perhaps it was keyed incorrectly or was a trick question. The latter seems to be confusing exam-takers and so they visit it many times. Thus, we are able to see more context around an item that might already have been flagged by CTT.

We do not propose that behavioural analysis is a replacement for CTT; knowing the number of item visits might not provide us with much insight without also knowing the Diff-I and Disc-I. Indeed, as has been pointed out by Chiavaroli2011, no single statistic should be used as an ultimate evaluation measure; rather they should be used to identify discrepant questions for further interpretation. Therefore, knowing that certain exam-taker behaviours are undoubtedly correlated with both the difficulty and discrimination of the items does provide us with another valuable tool in our kit for analysing exam quality.

Finally, the findings of the presented study have implications for practice. First, many studies suggest that using instructional interventions for test-taking enhancement can reduce anxiety and help students perform optimally in exams (eg, Hong et al., 2006). However, for instructors to provide reliable interventions, they first need to understand test-taking strategies (Bumbálková, 2021). The presented study provides evidence that students exhibit a wide range of behaviours during an exam, which should be considered a crucial step towards understanding exam-taking strategies. Second, artificial intelligence and LA are becoming more sophisticated in accurately identifying vulnerable students in need and providing actionable insight into students learning (Rosé et al., 2019). Successful examples include modelling students' engagement, knowledge state and learning tactics and strategies using trace data. Exam logs are no exception, meaning they can be used to trace exam-takers' actions to model their behaviours to help instructors identify students in need or exam items that need an instructor's attention (eg, poorly written items). The findings of the presented study (see Section 4.3) support the application of tools and systems that support instructors' oversight using exam logs where we could detect suspicious items that the instructor could make sense of.

CONCLUSION AND FUTURE WORK

The educational data revolution, empowered by the increasing use of technology in education, has enabled universities to collect rich data on their learners. In this paper, we examined how data collected via EAPs may contribute to item analysis. Reported results based on analysis of de-identified exam logs of 463 medical exam-takers suggest that log-based analysis can provide insights beyond what is captured by traditional methods of item analysis.

There are, however, several limitations in the presented study that restrict the generalisability of the results. One of the significant limitations is that the proposed research questions are answered based on data from a single course using the same type of questions and under one testing condition (ie, power test). It is important to acknowledge that exam-takers in different disciplines or institutes may have significantly different test-taking behaviour. In addition, whether the order of items (eg, easy to difficult or random ordering) can affect the

behaviours in this study is unknown. This is an important consideration in future research since the literature shows that item rearrangement might have an effect on item performance (Leary & Dorans, 1985). Future directions include replicating this study across different disciplines and institutes, taking into consideration the type of exam, examination condition and item rearrangement to evaluate the generalisability of our current findings.

There are several additional directions to pursue in future work. Further work that involves practitioners to evaluate the proposed approach against CTT can help bridge the gap between academia and practice in education. In particular, it would be interesting to evaluate whether practitioners can make sense of the features and find them to be explainable (Khosravi et al., 2022). If so, would they agree that they are valuable in complementing CTT for item analysis? Another interesting directions would be to explore whether we can use the logs generated by electronic assessment platforms towards analysing the behaviour of the exam-takers. This may assist in identifying the behaviours of successful exam-takers, which consequently may allow us to develop effective exam-taking strategies and personal recommendations for exam-takers on how to be testwise.

ACKNOWLEDGEMENTS

The first author would like to express his gratitude and acknowledgement to the government of Saudi Arabia, Ministry of Education, represented by Saudi Arabian Cultural Mission (SACM), for the scholarship and continued support. Open access publishing facilitated by The University of Queensland, as part of the Wiley - The University of Queensland agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

This research received no specific grant
from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

There is no conflict of interest in the work reported above.

DATA AVAILABILTY STATEMENT

Some of the data used in this research contain summative assessment marks. We are therefore unable to make this publicly available. However, an actual implementation in R alongside a sample data set representing our input data are released on GitHub at https://github.com/hlahza/BJET_beyond-item-analysis.

ETHICS STATEMENT

Ethical approval for conducting this research was received from the University of Queensland Human Research Ethics Committee # 2018000841.

ORCID

Hatim Lahza https://orcid.org/0000-0002-5849-9639
Hassan Khosravi https://orcid.org/0000-0001-8664-6117

ENDNOTE

[1] Ethical approval for conducting this research was received from The University of Queensland Human Research Ethics Committee (# 2018000841).

REFERENCES

Abdi, S., Khosravi, H., & Sadiq, S. (2020). Modelling learners in crowdsourcing educational systems. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (pp. 3–9). Springer International Publishing.

Abdi, S., Khosravi, H., & Sadiq, S. (2021). Modelling learners in adaptive educational systems: A multivariate glicko-based approach. In *Lak21: 11th international learning analytics and knowledge conference* (pp. 497–503). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448189

Abdi, S., Khosravi, H., Sadiq, S., & Darvishi, A. (2021). Open learner models for multi-activity educational systems. In *International conference on artificial intelligence in education* (pp. 11–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-22

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate elo-based learner model for adaptive educational systems. In *EDM: Proceedings of the 12th International Conference on Educational Data Mining* (pp. 228–233). International Educational Data Mining Society.

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, *39*(4), 821–824. https://doi.org/10.1177/001316447903900415

Barana, A., Conte, A., Fissore, C., Marchisio, M., & Rabellino, S. (2019, October). Learning analytics to improve formative assessment strategies. *Journal of e-Learning and Knowledge Society*, *15*, 75–88. https://www.je-lks.org/ojs/index.php/Je-LKSEN/article/view/1135057. https://doi.org/10.20368/1971-8829/1135057

Bauer, D., Kopp, V., & Fischer, M. R. (2007). Answer changing in multiple choice assessment change that answer when in doubt–and spread the word! *BMC Medical Education*, *7*(1), 1–5. https://doi.org/10.1186/1472-6920-7-28

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed–accuracy–revisits model. *Educational and Psychological Measurement*, *81*(2), 363–387. https://doi.org/10.1177/0013164420950556

Bumbálková, E. (2021). Test-taking strategies in second language receptive skills tests: A literature review. *International Journal of Instruction*, *14*(2), 647–664.

Chiavaroli, N., & Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of mcqs. *Bioscience Education*, *17*(1), 1–7. https://doi.org/10.3108/beej.17.8

Cleophas, C., Hoennige, C., Meisel, F., & Meyer, P. (2021). Who's cheating? Mining patterns of collusion from text and events in online exams. *INFORMS Transactions on Education*. Advance online publication. https://doi.org/10.1287/ited.2021.0260.

Costagliola, G., Fuccella, V., Giordano, M., & Polese, G. (2008). Monitoring online tests through data visualization. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 773–784.

Couchman, J. J., Miller, N. E., Zmuda, S. J., Feather, K., & Schwartzmeyer, T. (2016). The instinct fallacy: The metacognition of answering and revising during college exams. *Metacognition and Learning*, *11*(2), 171–185. https://doi.org/10.1007/s11409-015-9140-8

Cousin, G. (2006). *An introduction to threshold concepts* (Vol. 17, No. 1). Taylor & Francis.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eassessment: Amee guide no. 39. *Medical Teacher*, *31*(3), 192–206. https://doi.org/10.1080/01421590902792406

Dodonova, Y. A., & Dodonov, Y. S. (2012). Processing speed and intelligence as predictors of school achievement: Mediation or unique contribution? *Intelligence*, *40*(2), 163–171. https://doi.org/10.1016/j.intell.2012.01.003

Ellis, A. P. J., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology*, *33*(12), 2607–2629. https://doi.org/10.1111/j.1559-1816.2003.tb02783.x

Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics: Colloquium. *British Journal of Educational Technology*, *44*(4), 662–664. https://doi.org/10.1111/bjet.12028

Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology*, *10,*1131 https://doi.org/10.3389/fpsyg.2019.01131

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. *Computers in Human Behavior*, *134*, 107304.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best mcqs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, *62*(2), 142.

Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research*, *99*(3), 144–155. https://doi.org/10.3200/JOER.99.3.144-155

Ifenthaler, D., & Greiff, S. (2021). Leveraging learning analytics for assessment and feedback. In *Online learning analytics* (pp. 1–18). Auerbach Publications.

Ifenthaler, D., Greiff, S., & Gibson, D. (2018). Making use of data for assessments: Harnessing analytics and data science. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *Second handbook of information technology in primary and secondary education* (pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-319-53803-741–1

Jordan, S. (2013). E-assessment: Past, present and future. *New Directions*, *9*(1), 87–106. https://journals.le.ac.UK/ojs1/index.php/new-directions/article/view/504. https://doi.org/10.29311/ndtps.v0i9.504

Jung Kim, Y.-M. (2001). *Investigation of neel's new item analysis technique* (john h. neel) [doctoral dissertation] (AAI3023296). Georgia State University. https://www.proquest.com/docview/304694669/

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge. https://doi.org/10.4324/9781315708591-3

Karelia, B. N., Pillai, A., & Vegada, B. N. (2013). The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year ii mbbs students. *IeJSME*, *7*(2), 41–46.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2008). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007, August). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kupiainen, S., Vainikainen, M.-P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills. *Journal of Educational Psychology*, *106*(3), 627–638. https://doi.org/10.1037/a0035507

Lang, C., Wise, A., Siemens, G., & Gasevic, D. (2017). *Handbook of Learning Analytics*. SOLAR, Society for Learning Analytics and Research. https://doi.org/10.18608/hla17

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413.

Lee, Y. (2019). *Estimating student ability and problem difficulty using item response theory (irt) and trueskill*. Information Discovery and Delivery.

Livingston, S. A. (2006). Item analysis. In *Handbook of test development* (pp. 421–441). Routledge.

Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. *Computers Education*, *62*, 72–87. https://doi.org/10.1016/j.compedu.2012.10.021

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Pagni, S. E., Bak, A. G., Eisen, S. E., Murphy, J. L., Finkelman, M. D., & Kugel, G. (2017). The benefit of a switch: Answer-changing on multiple-choice exams by first-year dental students. *Journal of Dental Education*, *81*(1), 110–115. https://doi.org/10.1002/j.0022-0337.2017.81.1.tb06253.x

Palmiero, C., & Cecconi, L. (2019). Use of learning analytics in formative and summative evaluation. *Journal of e-Learning and Knowledge Society*, 15(3), 89–99 . https://doi.org/10.20368/1971-8829/1135019

Papamitsiou, Z., & Economides, A. A. (2017). Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us? *Computers in Human Behavior*, *75*, 423–438. https://doi.org/10.1016/j.chb.2017.05.036

Papamitsiou, Z., & Economides, A. (2014). Students' perception of performance vs. actual performance during computer-based testing: A temporal approach. In *Inted2014 proceedings* (pp. 401–411). IATED.

Papamitsiou, Z., & Economides, A. A. (2015). Temporal learning analytics visualizations for increasing awareness during assessment. *International Journal of Educational Technology in Higher Education*, *12*(3), 129–147. https://doi.org/10.7238/rusc.v12i3.2519

Papamitsiou, Z., Economides, A. A., Pappas, I. O., & Giannakos, M. N. (2018). Explaining learning performance using response-time, self-regulation and satisfaction from content: An fsQCA approach. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 181–190). Association for Computing Machinery. https://doi.org/10.1145/3170358.3170397

Papamitsiou, Z. K., & Economides, A. A. (2013). Towards the alignment of computer-based assessment outcome with learning goals: The LAERS architecture. In *2013 IEEE Conference on e-Learning, e-Management*

and e-Services (pp. 13–17). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/IC3e.2013.6735958

Papamitsiou, Z. K., Terzis, V., & Economides, A. A. (2014). Temporal learning analytics for computer based testing. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 31–35). Association for Computing Machinery. doi. https://doi.org/10.1145/2567574.2567609

Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, *50*(6), 2943–2958.

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 480–489). Association for Computing Machinery. https://doi.org/10.1145/3375462.3375498

Sim, S.-M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore*, *35*(2), 67.

Stenlund, T., Eklöf, H., & Lyrén, P.-E. (2017). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 4–20. https://doi.org/10.1080/0969594X.2016.1142935

Stenlund, T., Lyrén, P.-E., & Eklöf, H. (2018). The successful test taker: Exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education*, *33*(2), 403–417. https://doi.org/10.1007/s10212-017-0332

Thillmann, H., Gößling, J., Marschner, J., Wirth, J., & Leutner, D. (2013). Metacognitive knowledge about and metacognitive regulation of strategy use in self-regulated scientific discovery learning: New methods of assessment in computer-based learning environments. In *International handbook of metacognition and learning technologies* (pp. 575–588). Springer. https://doi.org/10.1007/978-1-4419-5546-3

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, *4,—*. https://doi.org/10.3389/feduc.2019.00049

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. https://doi.org/10.3102/1076998607302626

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064–1070.

Wauters, K., Desmet, P., & Van Noortgate, W. (2010). Monitoring learners' proficiency: Weight adaptation in the elo rating system. In *Proceedings of the 4th international conference on educational data mining 2011*(pp.247-*)*. International Educational Data Mining Society.

Wiggins, B. C. (2000). *Detecting and dealing with outliers in univariate and multivariate contexts*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY, November, 15–17, 2000.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252. https://doi.org/10.1080/08957347.2015.1042155

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lahza, H., Smith, T. G., & Khosravi, H. (2022). Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. *British Journal of Educational Technology*, *00*, 1–20. https://doi.org/10.1111/bjet.13270