International Journal of Artificial Intelligence in Education (2019) 29:484–506
https://doi.org/10.1007/s40593-019-00183-1

ARTICLE

Are MOOC Learning Analytics Results Trustworthy? With Fake Learners, They Might Not Be!

Giora Alexandron[1] · Lisa Y. Yoo[2] · José A. Ruipérez-Valiente[2] · Sunbok Lee[3] · David E. Pritchard[2]

Published online: 2 July 2019

Abstract

The rich data that Massive Open Online Courses (MOOCs) platforms collect on the behavior of millions of users provide a unique opportunity to study human learning and to develop data-driven methods that can address the needs of individual learners. This type of research falls into the emerging field of *learning analytics*. However, learning analytics research tends to ignore the issue of the reliability of results that are based on MOOCs data, which is typically noisy and generated by a largely anonymous crowd of learners. This paper provides evidence that learning analytics in MOOCs can be significantly biased by users who abuse the anonymity and open-nature of MOOCs, for example by setting up multiple accounts, due to their amount and aberrant behavior. We identify these users, denoted *fake learners*, using dedicated algorithms. The methodology for measuring the bias caused by fake learners' activity combines the ideas of Replication Research and Sensitivity Analysis. We replicate two highly-cited learning analytics studies *with* and *without* fake learners data, and compare the results. While in one study, the results were relatively stable against fake learners, in the other, removing the fake learners' data significantly changed the results. These findings raise concerns regarding the reliability of learning analytics in MOOCs, and highlight the need to develop more robust, generalizable and verifiable research methods.

Keywords Learning Analytics · MOOCs · Replication research · Sensitivity analysis · Fake learners

✉ Giora Alexandron
giora.alexandron@weizmann.ac.il

Extended author information available on the last page of the article.

Preface: The Beginning of this Research

During 2015 we were working on recommendation algorithms in MOOCs. However, we ran into a strange phenomenon – the most successful learners seemed to have very little interest in the course materials (explanation pages, videos), and they mainly concentrated on solving assessment items. As a result, the recommendation algorithms sometimes recommended skipping resources that we thought should be very useful. The hypothesis that these are learners who already know the material (e.g., Physics teachers) did not match the demographic data that we had on these users.

One day, we received a strange email from one of the users. The user complained about a certain question, claiming that a certain response that was correct a week ago is now rejected by the system as incorrect. Since it was a parameterized question (randomized per user), we suspected that the user viewed it from two different accounts. Connecting this with the strange pattern of users who achieved high-performance without using the resources, we realized that we bumped into a large-scale phenomenon of Copying Using Multiple Accounts (CUMA) (Alexandron et al. 2015a, 2017; Ruipérez-Valiente et al. 2016).

Our research on detecting and preventing CUMA started as a spin-off of the unexplained bias in predictive modeling. Three years later, we return to investigate the bias issue and its effect on learning analytics results.

Introduction

Modern digital learning environments collect rich data that can be used to improve the design of these environments, and to develop 'intelligent' mechanisms to address the needs of learners, instructors, and content developers (Siemens 2013; U.S. Department of Education 2012). MOOCs, which collect fine-grained data on the behavior of millions of learners, provide "unparalleled opportunities to perform data mining and learning experiments" (Champaign et al. 2014) (p. 1). A partial list of studies includes comparing active vs. passive learning (Koedinger et al. 2015), how students use videos (Kim et al. 2014a, b; Chen et al. 2016), which instructional materials are helpful (Alexandron et al. 2015b; MacHardy and Pardos 2015), recommending content to learners on real-time (Pardos et al. 2017; Rosen et al. 2017), providing analytics to instructors (Ruipérez-Valiente et al. 2017c), predicting drop-out (Xing et al. 2016), to list a few. These studies fall into the emerging disciplines of Learning Analytics (LA), Educational Data Mining (EDM), and Artificial Intelligence in Education (AIED). A recent review of the existing literature on the application of data science methods to MOOCs can be found in Romero and Ventura (2017). While 'Big Data in Education' is mostly associated with MOOCs and other learning at scale applications, it is also very relevant to other widely distributed platforms, such as Moodle (Luna et al. 2017).

Such research uses data-intensive methods that draw on machine learning, data mining, and artificial intelligence. These methods seek to extract meaningful structures from the data, which can have predictive value in inferring the future behaviors of students (Qiu et al. 2016). Thus, their reliability is highly determined by the

quality of the data. Yudelson et al. (2014) showed that models fitted on high quality data can outperform models that are fitted on a larger dataset that is of lower quality – a sort of the 'ed-tech variant' of Peter Norvig's "More data beats clever algorithms, but better data beats more data".[1]

One of the main issues that can affect the quality of the data is noise, which is anything that is not the 'true' signal (Silver 2012). Noise that appears as outliers can be identified and removed relatively easy using outlier detection methodologies (Hodge and Austin 2004). However, by definition, when there are too many 'outliers' in a certain direction, they are not 'outliers' anymore, rendering outlier detection methods ineffective. The same holds for data coming from complex, semi-structured or unstructured domains, such as MOOCs, in which a plethora of behaviors are expected (Qiu et al. 2016). In such cases, there can be numerous outliers in various directions that are caused by genuine learning activity. Eventually, unfiltered noise can significantly affect various analytics computed on the data. For example, Du et al. (2018) discussed how learning analytics perform differently on specific subgroups in MOOCs, and proposed a methodology to discover such subgroups, which is based on Exceptional Model Mining (EMM). The focus of our paper is *not* on discovering subgroups, but on showing that a certain subgroup (fake learners) has exceptional behavior that can bias analytics. In the future, it can be interesting to check if EMM techniques can identify subgroups of users who are actually fake learners.

In addition to the issue of noise, there is the issue of prior assumptions on the model. Educational data mining methods make such assumptions on the process that generated the data, either for choosing the objectives to optimize as proxies of the 'true' goal, for feature engineering, or for data abstraction (Perez et al. 2017). Thus, making wrong assumptions on the process can largely affect the validity of the results. Being able to model various subgroups is crucial for designing tailored pedagogic interventions (Kiernan et al. 2001). This is especially true for the diverse population of the global classroom – MOOCs. We note that making no 'modeling' assumptions and relying on Big Data alone, instead of as a supplement to traditional data analysis methodologies ('Big Data Hubris'), can also lead to large errors, as in the case of Google Flu Trends (Lazer et al. 2014).

Not only that the accuracy of machine learning (ML) models is highly affected by such issues, validating these models is a scientific and technological challenge (Seshia and Sadigh 2016). ML models are mathematical functions that are learned from the data using statistical learning methods (Hastie et al. 2001). Due to their probabilistic nature and complex internal structure, it is difficult to question their reasoning (Krause et al. 2016). This is even more the case when such models are encapsulated within artificial intelligence or analytics solutions that use them as 'black-boxes' for automatic or semi-automatic decision making. This is a major concern of the Big Data era (Müller et al. 2016), which is highly relevant to the educational domain as well (Pardo et al. 2016; O'Neil 2017).

[1] https://www.azquotes.com/author/49745-Peter_Norvig

Fake Learners Typically, data-driven education research makes the implicit assumption that the data that are used represent genuine learning behavior. However, recent studies revealed that MOOCs can contain large groups of users who abuse the system in order to receive certificates with less effort (Alexandron et al. 2015a, 2017; Ruipérez-Valiente et al. 2016, 2017a; Northcutt et al. 2016). As these users do not rely on 'learning' to achieve high performance, we denote these users as *fake learners* (and use the term *true learners* to describe the genuine ones).

Fake learners introduce noise to statistical models that seek to make sense of learning data. The effect of this noise depends mainly on the amount of fake learners, and on how aberrant their behavior is. In order to remove their effect by filtering them out from the data, they must first be detected. However, this typically involves sophisticated algorithms that are currently not available as 'off-the-shelf' tools that can be used to clean the data. Also, while this study relies on algorithms for detecting two types of fake learning methods – Copying Using Multiple Accounts (CUMA) and *unauthorized collaboration* (Ruipérez-Valiente et al. 2017a), there might be more types of fake learners who currently sneak under the radar.

The risk that the aberrant behavior exhibited by CUMA users, and their prevalence ($>$ 10% of certificate earners), which can bias learning analytics results, was raised in our previous work (Alexandron et al. 2017), but remained an open question. The goal of the current research is to address the bias issue directly. In this paper we demonstrate for the first time, to the best of our knowledge, that fake learners can significantly affect learning analytics results. We also provide a sort of 'future projection' to what will happen if the number of fake learners increases (a likely consequence of MOOC certificates gaining more value due to the current transition to MOOC-based degrees Reich and Ruipérez-Valiente 2019). The findings that we report here significantly extend a preliminary report from this work (Alexandron et al. 2018).

Research Questions We study the following Research Questions (RQs):

1. (RQ1) Is there a significant difference between the 'fake' and 'true' learners with respect to various performance measures, and to the amount of use of the course instructional materials?
2. (RQ2) Can this difference bias the results of learning analytics models in a significant way?

Replication Research Our research approach combines the ideas of Sensitivity Analysis (Saltelli et al. 2000) and Replication Research (Open Science Collaboration 2015). We pick two highly-cited learning analytics MOOC studies, and evaluate how sensitive to fake learners are findings obtained with a similar methodology on new data from our MOOC. This is done by replicating each of the studies *with* and *without* fake learners' data, and comparing the results. This is not a full, but a partial replication. Specifically, we focus on the parts that study the characteristics of effective learning behaviors.

The studies that we replicate are "Correlating Skill and Improvement in 2 MOOCs with a Student's Time on Task" (Champaign et al. 2014), and "Learning is Not

a Spectator Sport: Doing is Better than Watching for Learning from a MOOC" (Koedinger et al. 2015). They appeared on the First (2014) and the Second (2015) ACM Conference on Learning@Scale,[2] which is a premier venue for interdisciplinary research at the intersection of the learning sciences and computer science, with specific focus on large scale learning environments such as MOOCs.

Contribution This study provides the first evidence, to the best of our knowledge, that learning analytics research in MOOCs can be significantly biased by the aberrant behavior of users who abuse the open nature of MOOCs. This issue raises concerns regarding the reliability of learning analytics in MOOCs, and calls for more robust, generalizable, and verifiable research methods. A systematic approach for addressing this issue, within the conceptual framework of Educational Open Science (van der Zee and Reich 2018), is large-scale replication research. A significant stride towards making such research technologically feasible is made by the MOOC Replication Framework (MORF) (Gardner et al. 2018).

Materials and Methods

In this section we describe the experimental setup, the data, and the data mining algorithms. Some of the methodological contents of this section have been reused from Ruipérez-Valiente et al. (2016) and Alexandron et al. (2017)

Experimental Setup

The Course The context of this research is MITx Introductory Physics MOOC 8.MReVx, offered on edX in Summer 2014.[3] The course covers the standard topics of a college introductory mechanics course with an emphasis on problem solving and concept interrelation. It consists of 12 mandatory and 2 optional weekly units. A typical unit contains three sections: instructional e-text/video pages (with interspersed concept questions, also known as checkpoints), homework, and a quiz. Altogether, the course contains 273 e-text pages, 69 videos, and about 1000 problems.

Research Population The research population consists of 478 certificate earners, out of the 13,500 users who registered for the course. Overall, 502 users earned a certificate, but we removed beta-testers, users who help validate content before the course is published, from the analysis. Gender distribution was 83% males, 17% females. Education distribution was 37.7% secondary or less, 34.5% College Degree, and 24.9% Advanced Degree. Geographic distribution includes US (27% of participants), India (18%), UK (3.6%), Brazil (2.8%), and others (total of 152 countries).

Data The data for this study consists of learners' clickstream data, which mainly include video events (play, pause, etc.), responses to assessment items, and navigation

[2] https://learningatscale.acm.org/

[3] https://courses.edx.org/courses/MITx/8.MReVx/2T2014/course

to course pages, yielding about half a million data points. In addition, we use the course structure files, which hold information that describes the course elements and the relations between them (e.g., the page in which a question resides).

Fake Learners: Definition and Detection

We define *fake learners* as users who apply unauthorized methods to improve their grade. This definition emphasizes the fact that the apparent behavior of fake learners does not explain significant aspects of their performance (can be achieved without 'learning', at least of Physics, in the case of our course), that it is systematic, and goal-oriented (as opposed to 'gaming the system' (Baker et al. 2008), for example). In a few occasions we also use the term 'cheating', but as a general issue, and we deliberately avoid referring to 'fake learners' as 'cheaters'. This is because the question of what should be regarded as 'cheating' in MOOCs is an issue that requires a discussion that is out of the scope of this paper, going well beyond the technical question of whether the user broke the edX "Terms of Service & Honor Code".[4]

Currently, we have means to identify two types of such methods:

Copying Using Multiple Accounts (CUMA) This refers to users who maintain multiple accounts. A *master* account that receives credit, and *harvesting* account/s that are used to collect the correct answers (typically by relying on the fact that many questions provide the full answer, or at least true/false feedback, after exhausting the maximum number of attempts) (Alexandron et al. 2015a; Ruipérez-Valiente et al. 2016). We note that in this method the multiple accounts are used by the *same* person. We use the algorithm described in Alexandron et al. (2017). The algorithm detects 65 master accounts out of the 478 certificate earners (as noted in Alexandron et al. 2017, the algorithm is designed to provide a *lower bound* on the true number of *master* accounts). Among masters and harvesters, only the master accounts are considered as fake learners (the harvesting accounts are not certified).

Collaborators This definition refers to MOOC learners who collaborate with peers and submit a significant portion of their assignments together. This is explicitly forbidden by the edX Honor Code, "unless collaboration on an assignment is explicitly permitted", which was not the case. To detect such collaboration, we use the algorithm of Ruipérez-Valiente et al. (2017a),[5] which uses dissimilarity metrics to find accounts that tend to submit their assignments in close proximity in time. Overall, the algorithm identifies 20 (∼4%) of the certificate earners as submitting a significant portion of their assignments with peers.

As there are users who use both methods, we give the CUMA algorithm priority when conducting analyses that require to assign a user to one of the groups ('CUMA Users' or 'Collaborators'), as it represents a more specific behavioral

[4] https://www.edx.org/edx-terms-service

[5] source code: https://github.com/jruiperezv/close_submitters_algorithm

pattern. Among the unauthorized collaborators, 11 also used CUMA. Hereafter we refer as 'collaborators' to the 9 accounts who were not CUMA users.

Measures of Learners' Performance

To date there is no standard and well accepted method to evaluate the performance of MOOC learners. In the MOOC literature, the most common measures are most likely *grade*, and the binary yes/no for certification. We use these, in addition to more robust methods that draw on Psychometrics and Item Response Theory (IRT) (Meyer and Zhu 2013; Champaign et al. 2014). The measures that we use are listed below.

Grade Total points earned in the course (60 points is the threshold for certification). The main issue with this measure is that it is very sensitive to *which* and *how many* items a learner attempts. In addition, it does not consider the attempt in which the learner succeeded (most items allow multiple attempts). Also, the variability that we observed on this measure was very low. Due to these limitations, we find this measure less useful as a valid and reliable measurement. However it is the most common measure, and what edX instructors receive.

Proportion Correct on First Attempt (CFA) The proportion of items, among the items that the student attempted, that were answered correctly on the first attempt. While CFA is a simple and straightforward approximation of students performance, which in this MOOC is highly correlated with more robust measures (e.g, IRT), it is also very sensitive to *which* items a learner chooses to attempt (choosing easy items will lead to higher CFA).

Ability Student's ability using a 2PL IRT model. Model is fitted on the first attempt matrix of the certificated users *(N=502)*, and item set that contains questions attempted by at least 50% of these users. We chose IRT because students' IRT ability scores are known to be independent of the problem sets each student tried to solve (De Ayala 2009). Missing items are treated using mean imputation (Donders et al. 2006). The model is fitted on a standard laptop using R's *TAM* package.[6]

Weekly Improvement Per student, this is interpreted as the slope of the regression line fitted to the weekly IRT ability measure (namely, the result of fitting 2PL IRT on each week of the course in separate) (Champaign et al. 2014). One of the important issues that must be addressed during the calculation of the IRT slopes is to set up the common scale across weekly IRT scores. IRT is a latent variable model, and a latent variable does not have any inherent scale. Therefore, each IRT estimation defines its own scale for the latent variable. Equating is the process of transforming a set of scores from one scale to another. We used mean and sigma equating to set up a common scale across weekly IRT scores. The equated IRT slope captures *the change*

[6] https://cran.r-project.org/web/packages/TAM/TAM.pdf

in students' relative performance during the course. For example, a student who has average performance in all the weeks, will have 0 relative improvement.

Mean Time on Task The average time the student spent on an item. For multiple attempts, it is composed of the sum of time for all attempts. The time for each attempt is operationalized as the delta between the time of the attempt, and the time of the previous action (navigating into the page, submission to previous item on the same page, or previous submission to this item in case of multiple attempts; we do not accumulate durations over 15 minutes, assuming that the user disengaged from the system Champaign et al. 2014). Time on task, or response time, plays an important role in cognitive ability measurement (Goldhammer 2015). Kyllonen and Zu (2016) stated that "A recurring question has been whether speed and level are simply two interchangeable measures of the same underlying ability or whether they represent different constructs" (p. 2).

Results

The results are organized into three subsections. The first subsection provides descriptive statistics that demonstrates the differences between fake and true learners with respect to fundamental behavioral characteristics. The second and third subsections present the results of replicating Koedinger et al. (2015) and Champaign et al. (2014), respectively.

Differences in Behavioral Characteristics

Time on Course Resources First, we measure the amount of time that fake learners spent on different course resources, compared to true learners. We consider:

- Reading time: Time that the user spent on explanatory pages.
- Watching time: Time that the user spent on videos.
- Homework time: Time spent in pages that contain homework items.

Figure 1 presents the time that fake and true learners spent on each resource type. As can be seen, fake learners spent significantly less time on each type of the instructional resources. This is confirmed with a two-sided Mann-Whitney U test ($n_{fake} = 72$, $n_{true} = 406$, $p.value < 0.01$). For Watching Time, median values for fake and true learners are 0.3 and 1.6 hours, respectively ($U = 21,040$); For Reading Time, median values for fake and true learners are 7.2 and 15.9 hours, respectively ($U = 21,499$); For Homework Time, median values for fake and true learners are 7.7 and 16.0 hours, respectively ($U = 21,934$).

Proportion of Items Solved The proportion of assessment items that true and fake learners attempted (successfully or not) is another metric on which we compare the behavior of the groups. A main reason is that solving assessment items, especially

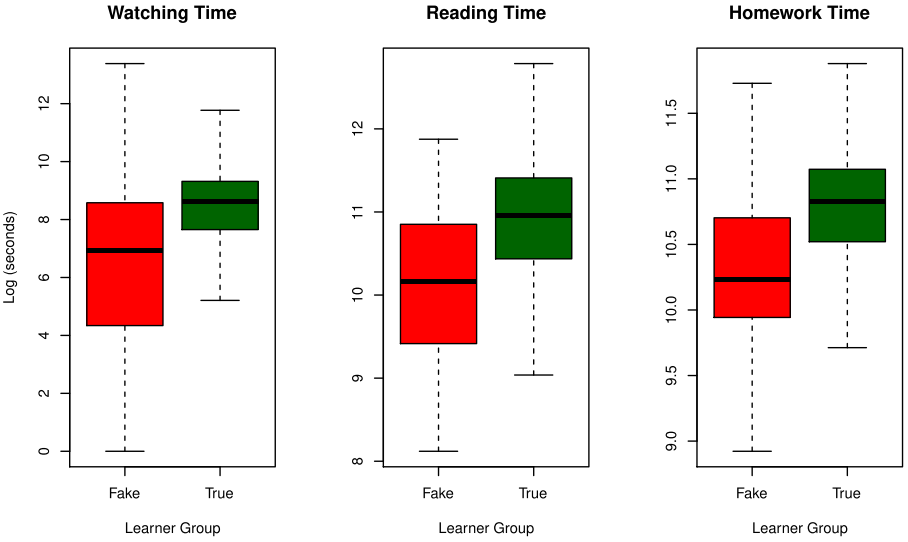


Fig. 1 Time on instructional resources

ones that do not contribute much to the final grade (available as formative assessment), and after the learner secured enough points to receive a certificate, is a clear indication of motivation to learn.

The course contains mainly three types of assessment items: Checkpoint, Homework, and Quiz (see Section "Experimental Setup"). They are analyzed separately because of their different characteristics with respect to weight (points for solving them), and the easiness of getting the correct answer without effort (e.g., whether the 'show answer' option is enabled for most of them after exhausting the possible attempts).

We assume that fake learners would factor that into their decision of whether to spend time on these items. For example, since Checkpoint items have low weight, we assume that fake learners would show less interest in solving them. Quiz items have high weight, but are harder to copy (no 'show answer', only true/false feedback). Homework offers relatively high weight and have 'show answer' enabled, which probably makes them ideal for fake learners (high 'return on investment').

Figure 2 presents the proportion of items solved by each group. As in the case of the time spent on resources (previous subsection), there is a clear difference between the groups, with fake learners trying less items. This is confirmed with a two-sided Mann-Whitney U test ($n_{fake} = 72$, $n_{true} = 406$, $p.value < 0.05$). For Checkpoint items, median values for fake and true learners are 0.69 and 0.75, respectively ($U = 18,442$); For Quiz items, median values for fake and true learners are 0.58 and 0.64, respectively ($U = 18,752$); For Homework items, median values for fake and true learners are 0.47 and 0.48, respectively ($U = 17,088$).

Interestingly, and as we suspected, the difference between the groups is smaller on Homework items which provide to fake learners the appealing combination of high

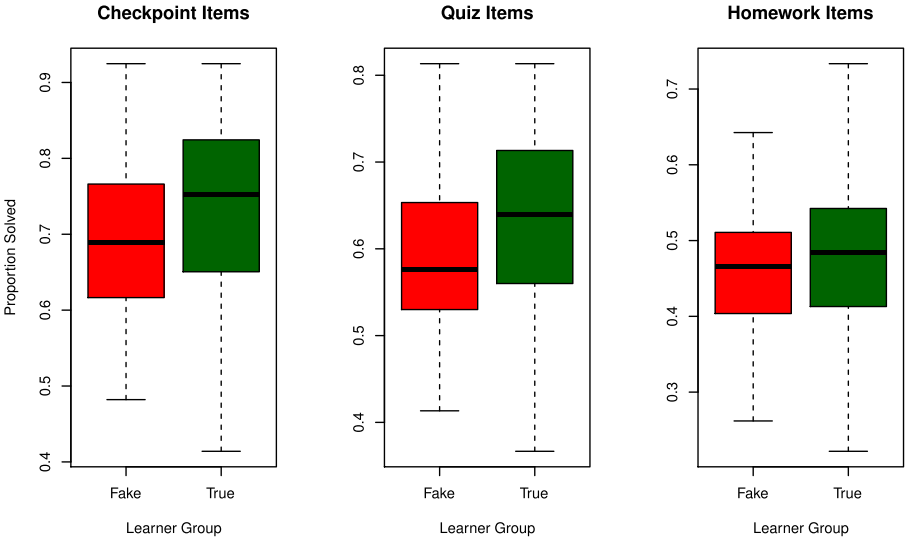


Fig. 2 Proportion of items solved by category

weight and a 'show answer' feature. This is furthered discussed in the "Discussion" section.

Performance Measures Figure 3 illustrates the differences between fake and true learners with respect to the measures of learners performance that were defined in the "Materials and Methods" section.

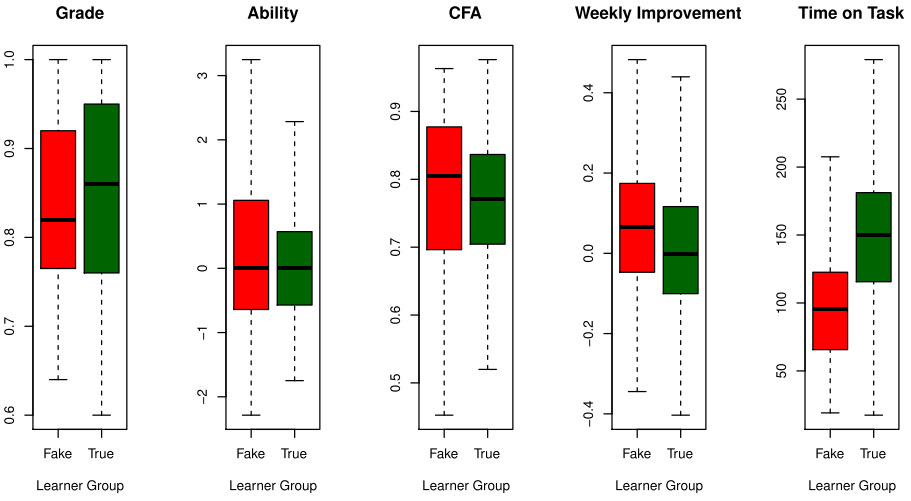


Fig. 3 Performance measures

Fake learners are significantly faster than true learners (median time-on-task is 97 vs. 150 seconds, respectively; $p - value < 0.001$). Another measure on which they are better than true learners is *weekly improvement*. This means that fake learners tend to improve (relative to the other learners) during the course. This is in line with the findings of Alexandron et al. (2017), which reported that the number of items that CUMA users copied tended to increase significantly as the course progressed.

On the other metrics (*grade*, *ability*, CFA) the two populations do not differ significantly (fake learners have a higher CFA, and lower *grade*, but with $p-value > 0.05$). However, we do find on these metrics a significant difference *within* the fake learners cohort, between the CUMA users and the collaborators. This is demonstrated in Fig. 3. CUMA users have higher *grade* (0.85 vs. 0.77), ability (0.21 vs. −0.66), and CFA (0.79 vs. 0.67), than collaborators, all with significant *p-values*.

Summary of the Differences Overall, we see that fake learners spent much less time on course resources, and attempted less items. In the case of response time, we see that fake learners solve exercises much faster. Regarding success metrics, fake learners have a higher *weekly improvement*. On the other success metrics (*grade*, *ability* and CFA), on average there is no significant difference between true and fake learners.

Replication Study 1

Next, we examine the effect of the differences in the behavioral metrics presented above on the findings reported in "Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC" (Koedinger et al. 2015). Specifically, we concentrated on the third research question (RQ) – "What variations in course feature use (watching videos, reading text, or doing activities) are most associated with learning? And can we infer causal relationships?".

The analysis for this RQ is presented in the section titled "Variation in Course Feature Use Predict Differences in Learning Outcomes" (starting on p. 116, left column). It has two parts – "Exploratory Data Analysis", and "Causal Analysis", which we refer to as Sections "Analysis 1A" and "Replicating Analysis 1B", respectively.

Analysis 1A

This analysis characterizes students on three behavioral dimensions by performing a median split on each of the metrics – amount of videos played, number of pages visited, and number of activities started. A learner who is on the upper half of each split is referred to as 'Watcher', 'Reader', and 'Doer', respectively. This split yields 8 (2^3) subgroups.

The subgroups are compared on 2 global performance measures – 'Quiz Total', and 'Final Exam'. The conclusions regarding the quizzes are that "Doers do well on the quizzes [...] even without being on the high half of reading or watching", "doing the activities may be sufficient to do well on the quizzes". Regarding the final exam, it is found that doing is most important ("a higher final exam score is more typical of those on the higher half of doing"), but that doing is furthered enhanced

by watching, reading, or both. Altogether, the title of this analysis is that "Doing, not Watching, Better Predicts Learning", which supports the phrase "Doing is Better than Watching" in the title of the paper.

Replicating Analysis 1A on 8.MReVx 2014

In order to conduct this analysis on the data of 8.MReVx, we need to make a few adjustments. The definitions of Watcher, Reader, and Doer remain the same. Doer is computed based on the amount of items started, but this raises some issue as the *grade* and the IRT ability – the outcome measures, are also based on the items solved. To make these measures as independent as possible, we base the definition of Doer on checkpoint items (items within the units). This is to reflect the nature of 'active learning by doing', which we interpret as the main idea behind the doing profile, while using items that their direct contribution to the *grade* and IRT ability is minor.

The second major decision to make was on what the global measure of learning should be. In the original paper, Quiz Total and Final Exam are used. In 8.MReVx there is no Final Exam (there is a post-test that was taken by a very small number of learners), and the grade is not sufficient for this purpose, as the variability of the grade is very low, and its distribution among the learner profiles is quite uniform. Thus, we use IRT as a global measure of performance in the course.

The results of the analysis, with and without fake learners, are presented in Fig. 4. The left figure demonstrates the analysis for **all** learners (including fake learners). Within this figure, the leftmost, red bar represents the ability of 'Doers who are neither Watchers nor Readers'. This bar is the *highest*, meaning that this is the *most successful* group of learners. It is in line with the finding of Koedinger et al. (2015) that Doers can do well without watching videos or reading explanations (the original figure from Koedinger et al. (2015) is presented in Appendix 3, Fig. 10a).

The right figure presents the results of the same analysis **without fake learners** (only true learners). As can be seen, the performance of 'Doers who are neither

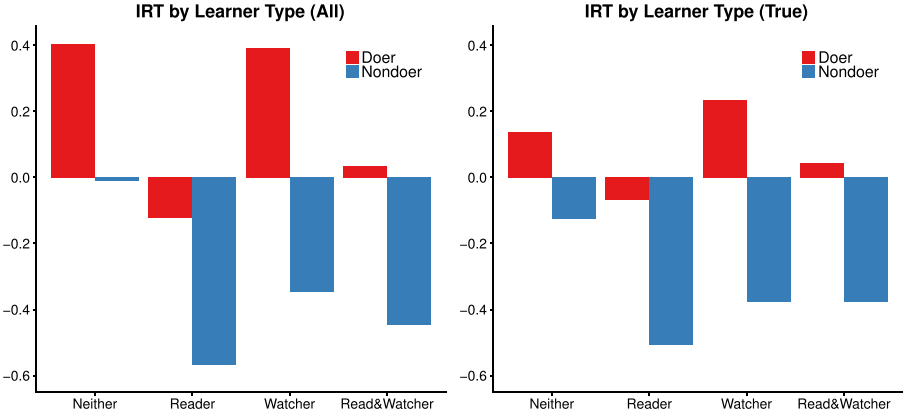


Fig. 4 IRT results: **a** All learners (including fake learners); **b** true learners only (fake learners removed)

Watchers nor Readers' drops sharply from mean IRT ability of 0.41 to mean IRT ability of 0.14. This change is statistically significant (using Bootstrap hypothesis testing; see below). **Without fake learners, 'Doers who are also Watchers' become the most successful group** (there is a small decrease in the performance of this group, which is statistically insignificant).

To verify that the effect is not an artifact we use Bootstrap hypothesis testing (MacKinnon 2009). Denote the group of all learners by L, and the size of the group of the 'true' learners by n. We estimate the 'Sampling distribution of mean IRT ability' of 'Doers who are neither Watchers nor Readers' and 'Doers who are also Watchers' using 1000 bootstrap samples of size n from L. The results show that the change to the mean IRT ability of 'Doers who are neither Watchers nor Readers' is statistically significant ($p - value < 0.05$), and that the change to the mean IRT ability of 'Doers who are also Watchers' is insignificant. A figure demonstrating the sampling distribution for both groups is provided in Appendix 1.

Doers still do better in all combinations, but the original conclusion that Doers can do well without watching videos or reading explanations becomes debatable when removing fake learners.

Replicating Analysis 1B

In the original analysis, Tetrad, a tool for causal inference, was used to evaluate whether associations between key variables – pre-test, use of course materials (doing, watching, reading), and outcomes (quiz total, final test), are potentially causal (the original graph from Koedinger et al. (2015) is presented in Appendix 3, Fig. 10b).

We replicated the same analysis using Tetrad on the data of 8.MReVx with and without fake learners. As a 'pre-test', we used the IRT score of the first week. The results are presented in Fig. 5. The left figure demonstrates the graph for **all** learners (including fake learners). The graph on the right is the result after removing fake learners, namely, for true learners only.

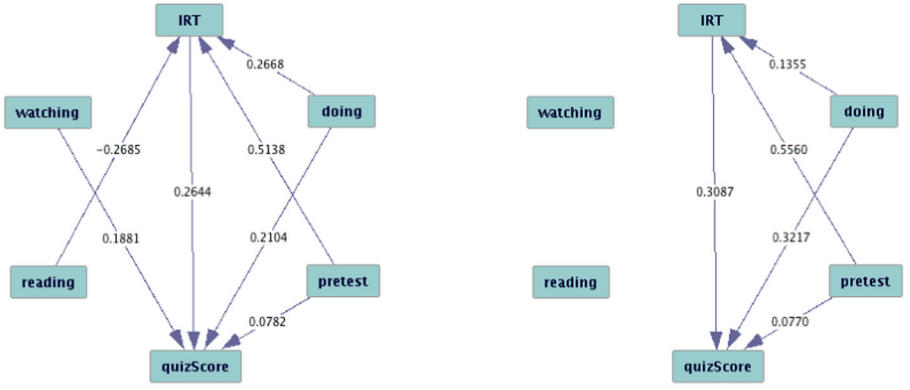


Fig. 5 Tetrad Results: **a** All learners (including fake learners); **b** true learners (fake learners removed)

As can be seen, the **causal graph changes significantly after removing fake learners data**. First, when removing fake learners (moving from the left to the right figures), two causal link disappear ($watching \rightarrow quizScore$ and $reading \rightarrow IRT$). Interestingly, the original link $reading \rightarrow IRT$ has a negative weight, meaning that *with* fake learners, reading is found to have a **negative effect** on IRT ability. The weights of the other edges change, but we do not interpret this as a qualitative difference. (We note that we could not use sampling techniques to evaluate the statistical significance of the change in the Tetrad results, as we did not find a way to run Tetrad's execution engine as a run-time library. The tool version that we use is the one that was used in the original paper (Koedinger et al. 2015), which is an old version that to the best of our knowledge does not support such usage.)

Future Projection

To evaluate the effect that an increase in the percentage of fake learners could potentially have on the analytics bias, we repeat the analyses of Sections "Analysis 1A" and "Replicating Analysis 1B" after increasing the amount of fake learners from $\sim$15% to $\sim$26%. This increase is achieved by simply duplicating the fake learners data, in order to maintain a similar multivariate distribution with respect to the behavioral characteristics that we measure.

The rationale for this 'simulation' analysis is twofold: First, the current $\sim$15% fake learners is a lower bound, and we assume that the actual amount of fake learners is higher. Second, it seems reasonable to assume that cheating in MOOCs will increase as the result of MOOC certificates gaining more value (Alexandron et al. 2017).

The effect on the bias is presented in Appendix 2. The effect on the bias of the Tetrad analysis is incremental (same edges, with slight change of weights; see Appendix 2, Fig. 9). The effect on *Analysis 1A* is significant, with "Doers who are neither Watchers nor Readers" becoming by far the most successful group (see Appendix 2, Fig. 8).

Summary of results - Replication Study 1

Based on the results of Sections "Analysis 1A" and "Replicating Analysis 1B", we conclude that the analysis of "What variations in course feature use are most associated with learning? And can we infer causal relationships?" – RQ3 from the paper (Koedinger et al. 2015) – changed in a meaningful way when replicated on the data of 8.MReVx with and without fake learners.

Replication Study 2

Another educational data mining study of the relation between which course materials learners use, and their success in the course, was presented in Champaign et al. (2014). The research objective is to understand the effectiveness of online learning materials, with the goal of improving the design of interactive learning environments. Among the "most striking features" that emerged from their analysis were (p. 18) "the large number of negative correlations between time spent on resource

use and skill level in 6.002x" and "the significant negative correlations between relative skill increase and time spent on any of the available instructional resources in 6.002x, accompanied by only one significant *positive* correlation" (the figure from the original paper is presented in Appendix 3, Fig. 11).

As in the case of Koedinger et al. (2015), what attracted our attention was the negative correlation between performance and use of (some of) instructional resources. Again, does this mean that the learning materials are unhelpful?

Thus, we replicate the analysis within Subsection "Correlations of Skill and Learning with Instructional Resource Use" (starts at p. 17). We note that the research found significant differences in the same correlations among two different MITx MOOCs (8.MReV 2013 and 6.002x 2012 – the first MOOC offered by MITx). Since these correlations seem to be course specific, our focus when replicating this analysis is not whether we receive the same results, but whether the results that we receive remain the same **with** and **without** fake learners.

The results are presented in Fig. 6 (the figure adopts the visualization used in Champaign et al. 2014). It shows the relation between the amount of time spent on various course resources, and certain performance metrics. For each pie, the outer circle is the whole group, and the inner is the same measure after removing fake learners

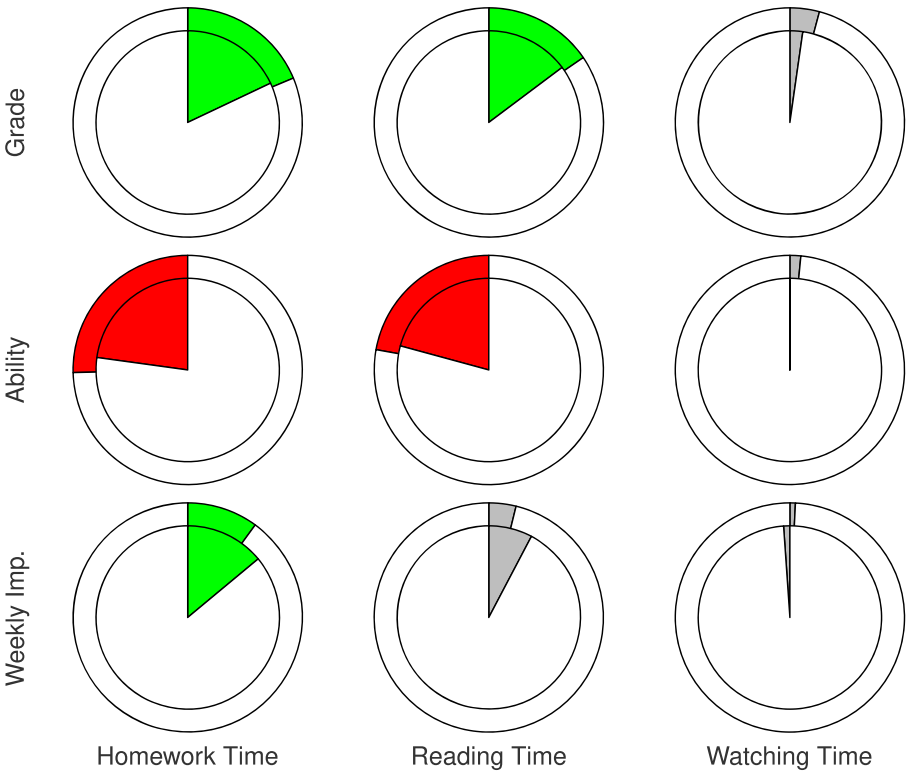


Fig. 6 Effect of fake learners on correlation between performance and time on course resources

from the data. The angle of the piece represents the size of the correlation. Clockwise angle represents positive correlation (colored with green), and counterclockwise represents negative correlation (colored in red). Gray color means $p - value > 0.05$. The difference between the angle of the outer circle, and the angle of the inner one, is the effect of fake learners' data on the correlation.

Let us examine the correlations with $p - value < 0.05$ (colored with red/green). With respect to Grade vs. Homework Time and Grade vs. Reading Time, there is almost no effect (angle of inner and outer piece is almost identical). With respect to Ability vs. Homework and Ability vs. Reading Time, we see a *negative* correlation, which is *reduced* when removing fake learners. With respect to weekly improvement vs. homework time, we see a *positive* correlation, which *increases* when removing fake learners.

Summary of results - Replication Study 2

For the three metrics that changed when removing fake learners' data – Ability vs. Homework Time, Ability vs. Reading Time, and Weekly Improvement vs. Homework Time – we see that after removing fake learners' data, the correlation moved in the positive direction. However, in each of the correlations we did not see a qualitative change, such as a negative correlation that becomes positive.

Discussion

Our results show that fake learners interact with the course materials in a very different way than true learners. For example, they attempt fewer questions and show minimal interest in the instructional materials. On the other hand, they exhibit high performance on various metrics – their IRT *ability*, *weekly improvement*, and CFA, are slightly higher, and their time-on-task is significantly faster. This is not surprising, as these users use means other than learning to achieve these results.

Due to their aberrant behavior, and their prevalence the MOOC that we study ($\sim$15% of the certificate earners), we suspected that fake learners can bias the results of learning analytics, especially those that deal with effective use of course resources. Our findings show that this is not an hypothetical risk. We use the methodology of Replication Research as means to focus our investigation on analytics that are acknowledged by the learning analytics research community as meaningful insights into what constitute effective learning behavior in MOOCs. Among the two studies that we replicated, one was relatively stable against fake learners, but on the other, the findings were biased in a significant way by fake learners' data.

Since the primary motivation for fake learning in 8.MReV is receiving a certificate with low effort (Alexandron et al. 2017), it is reasonable to assume that the same behavioral pattern of high performance and low resource use would characterize such learners in other MOOCs. Depending on the prevalence of fake learners among the learners, MOOC-based learning analytics research is vulnerable to bias by

fake learners. However, the amount of fake learners within MOOCs is still an open question.

The amount of CUMA in 8.MReV seems to generalize to other MOOCs (Alexandron et al. 2017), and we do not see reason to assume otherwise for collaborators. Obviously, there could be other fake learning methods, but we have not worked on new detection algorithms yet. Over time, we can expect that the percentage of fake learners, and subsequently their effect, may rise with the increase in the value of MOOC certificates, unless proper actions are taken against this. To evaluate the effect of increase in the number of fake learners, we conducted a simple 'simulation' analysis, illustrating the hypothetical bias caused by a 2X increase in the percentage of fake learners (see Section "Future Projection"). The results demonstrate that with 2X fake learners we can expect to see a significant increase in the bias.

Our research supplies some evidence that fake learners can lead to wrong inference on analytics aiming to address questions such as 'what are effective learning strategies?', 'which types of resources are helpful?', and so on. While 'global' correlations seem to be less vulnerable, studying selected cohorts like 'efficient learners' (e.g., ones who are fast and successful) would be very prone to bias due to fake learners' activity.

It is important to emphasize that the bias is due to the the fact that fake learners introduce noise into the data. Such noise can affect various types of computational models – for example, consider a recommendation engine that sequences content to learners in real-time. Such engines typically rely on machine learning models that are fitted to learners' data. Biased data can lead to modeling 'noise' instead of real predictor-outcome relationships. Since such machine learning models are typically encapsulated within 'policy' layers that use 'business' (in this case, pedagogy) logic to translate prediction into action, validating the recommendations becomes an extremely difficult task (Krause et al. 2016). As a thought experiment, imagine two competing MOOC content recommendation engines: an adversary engine that sends learners to random pages that they have not seen, and a Zone of Proximal Development (ZPD) engine that is tuned to challenge learners while keeping them within the ZPD. Now, assume that we have a sequence of pages that both engines recommended during a 15-minutes activity. What is the chance that an expert would identify which one is the adversary, and which one is the ZPD, without knowing the nitty-gritty of the ZPD engine?

Fake Accounts in Social Networks Malicious use of fake accounts is a common issue in social networks. Above all, the Facebook–Cambridge Analytica data scandal[7] has brought to public attention the issue of fake accounts, and how they can be used in malicious ways and on large-scale to collect data and affect social trends. Partly as the

[7] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

result of this 'wake-up call', Twitter recently announced that it has shut down 70 million fake/suspicious accounts since May 2018.[8] This (negative) similarity between MOOCs and social networks sheds some light on the fact that MOOCs are a learning environment which is also a global social platform.

Limitations The main limitations of our research are the fact that it is based on data from one MOOC, and on detection algorithms that detect only a subset of the fake learners in the course. Future research can expand this analysis to multiple MOOCs, and hopefully until then there will be algorithms for detecting other fake learning methods. In addition, it would be interesting to evaluate the effect on a wider set of learning analyses and machine learning models.

Summary and Conclusions

This study follows Replication Research methodology, and uses Sensitivity Analysis techniques to study how learning analytics can be biased by noisy data that include a significant amount of fake learners – learners who use illegitimate techniques to improve their grade. These users exhibit learning behaviors that are very different from those of 'true' learners, and achieve high performance. This can bias the analytics towards falsely identifying non-learning behaviors as effective learning strategies.

Our findings provide the first evidence, to the authors' knowledge, of how non-learning behaviors that are not modeled can significantly bias learning analytics results. The findings also point to the fact that cheating in educational settings can have consequences that go way beyond the issue of academic dishonesty.

To date, the issue of the reliability of learning analytics has received little attention within the learning analytics research community. As a first step, it is important to acknowledge that this is a real concern. Conveying this message is one of the major goals of this paper.

In order to address this issue, it is important to adopt more robust techniques for evaluating and validating learning analytics research, for example by encouraging and facilitating replication research at scale, and by developing advanced verification techniques. Another direction to take from this research is to develop detection methods that can generalize across platforms and course designs, e.g. by using ML (Ruipérez-Valiente et al. 2017b) or anomaly detection techniques (Alexandron et al. 2019).

In the verification domain, much can be learned from the hardware verification industry, which makes extensive use of sophisticated simulation methods, and from recent developments in the area of verification of autonomous vehicles, which deal with verifying complicated artificial intelligence systems.

Acknowledgments GA's research is supported by the Israeli Ministry of Science and Technology under project no. 713257.

[8] https://www.nytimes.com/2018/07/11/technology/twitter-fake-followers.html

Appendix 1: Sampling Distribution of Mean IRT Ability

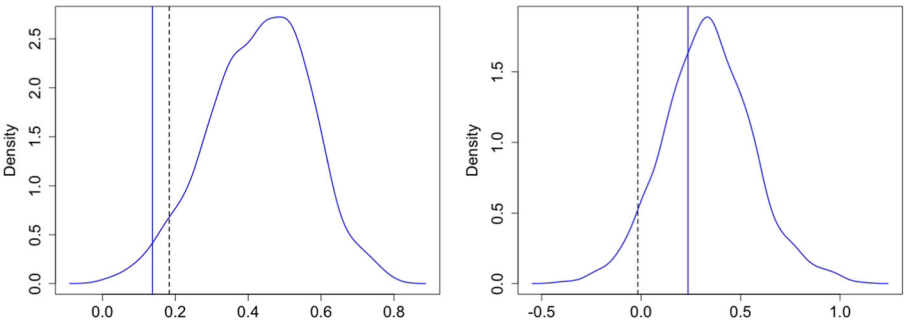


Fig. 7 Sampling Distribution of Mean IRT Ability for **a** Doers who are neither Watchers nor Readers **b** Doers who are also Watchers. The dashed vertical lines mark the 95% confidence interval, and the vertical blue lines mark the mean value without fake learners

Appendix 2: Replication Study 1 with 2X Simulated Fake Learners

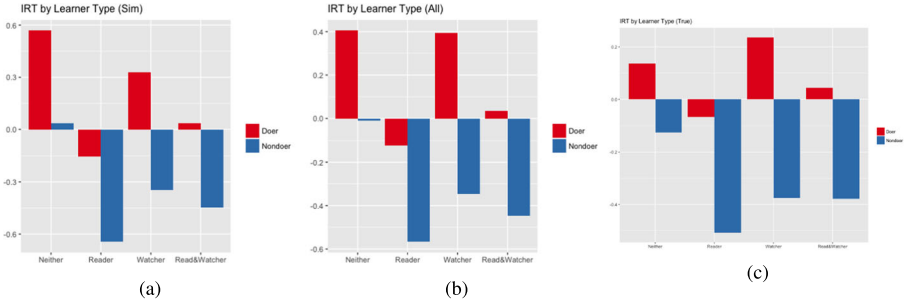


Fig. 8 IRT Results **a** With simulated 2x fake learners **b** Original data (All learners); **c** true learners (fake learners removed)0

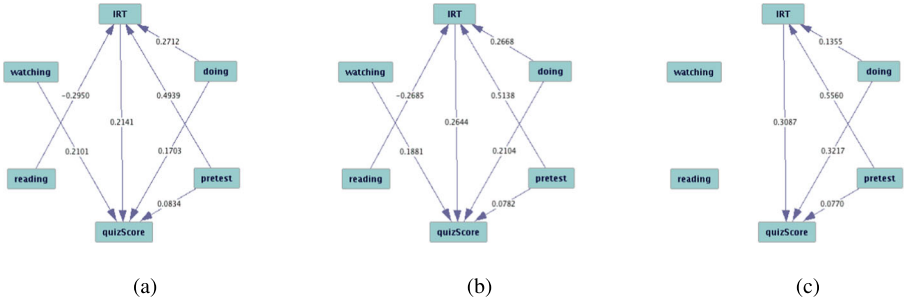
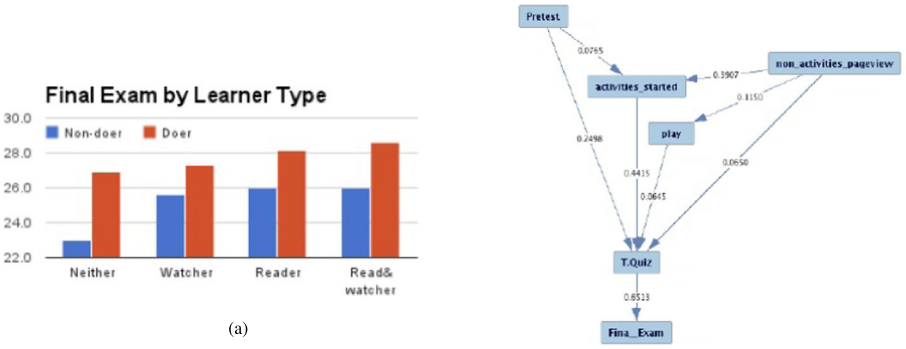


Fig. 9 Tetrad Results **a** With simulated 2x fake learners **b** Original data (all learners); **c** true learners (fake learners removed)

Appendix 3: Figures from Original Papers



(a) (b)

Fig. 10 Original figures from Koedinger et al. (2015) **a** Final grade by learner type **b** Causal model generated by Tetrad

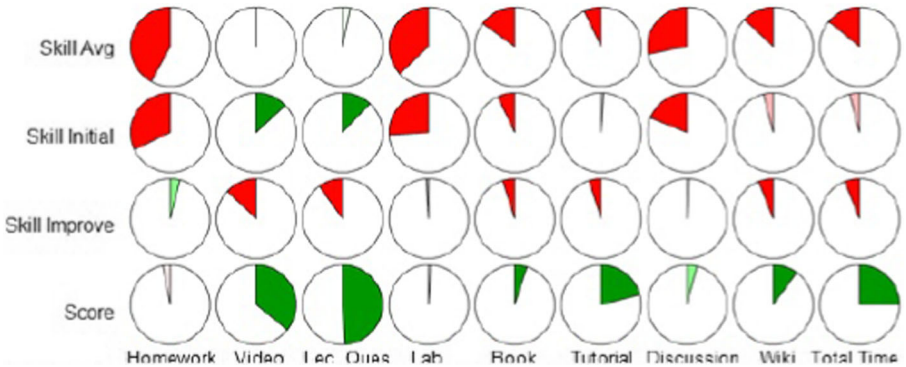


Fig. 11 Original figure from Champaign et al. (2014)

References

Alexandron, G., Ruipérez-Valiente, J.A., Pritchard, D.E. (2015a). Evidence of MOOC students using multiple accounts to harvest correct answers. Learning with MOOCs II, 2015.

Alexandron, G., Zhou, Q., Pritchard, D. (2015b). Discovering the pedagogical resources that assist students in answering questions correctly – a machine learning approach. In *Proceedings of the 8th international conference on educational data mining* (pp. 520–523).

Alexandron, G., Ruipérez-Valiente, J.A., Chen, Z., Muñoz-Merino, P.J., Pritchard, D.E. (2017). Copying@Scale using harvesting accounts for collecting correct answers in a MOOC. *Communication Education*, *108*, 96–114.

Alexandron, G., Ruipérez-Valiente, J.A., Lee, S., Pritchard, D.E. (2018). Evaluating the robustness of learning analytics results against fake learners. In *Proceedings of the thirteenth European conference on technology enhanced learning*: Springer.

Alexandron, G., Ruipérez-Valiente, J.A., Pritchard, D.E. (2019). Towards a general purpose anomaly detection method to identify cheaters in massive open online courses. In *Proceedings of the 12th international conference on educational data mining*.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 162–182.

Champaign, J., Colvin, K.F., Liu, A., Fredericks, C., Seaton, D., Pritchard, D.E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14, (March): 11–20*.

Chen, Z., Chudzicki, C., Palumbo, D., Alexandron, G., Choi, Y.-J., Zhou, Q., Pritchard, D.E. (2016). Researching for better instructional methods using AB experiments in MOOCs: results and challenges. *Research and Practice in Technology Enhanced Learning*, *11*(1), 9.

De Ayala, R. (2009). The theory and practice of item response theory. Methodology in the social sciences. Guilford Publications.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091.

Du, X., Duivesteijn, W., Klabbers, M., Pechenizkiy, M. (2018). Elba: exceptional learning behavior analysis. In *Educational data mining* (pp. 312–318).

Gardner, J., Brooks, C., Andres, J.M.L., Baker, R. (2018). Morf: a framework for MOOC predictive modeling and replication at scale. arXiv:1801.05236.

Goldhammer, F. (2015). Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 133–164.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The elements of statistical learning. Springer series in statistics*. New York: Springer.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Kiernan, M., Kraemer, H.C., Winkleby, M.A., King, A.C., Taylor, C.B. (2001). Do logistic regression and signal detection identify different subgroups at risk? implications for the design of tailored interventions. *Psychological Methods*, *6*(1), 35.

Kim, J., Guo, P.J., Cai, C.J., Li, S.-W.D., Gajos, K.Z., Miller, R.C. (2014a). Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on user interface software and technology - UIST'14* (pp. 563–572).

Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C. (2014b). Understanding in-video dropouts and interaction peaks in online lecture videos.

Koedinger, K.R., Mclaughlin, E.A., Kim, J., Jia, J.Z., Bier, N.L. (2015). Learning is not a spectator sport doing is better than watching for learning from a MOOC, pp. 111–120.

Krause, J., Perer, A., Ng, K. (2016). Interacting with predictions: visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5686–5697): ACM.

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14.

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Luna, J.M., Castro, C., Romero, C. (2017). Mdm tool: a data mining framework integrated into moodle. *Computer Applications in Engineering Education*, *25*(1), 90–102.

MacHardy, Z., & Pardos, Z.A. (2015). Toward the evaluation of educational videos using bayesian knowledge tracing and big data. In *Proceedings of the second (2015) ACM conference on learning @ scale, L@S '15* (pp. 347–350): ACM.

MacKinnon, J.G. (2009). Bootstrap hypothesis testing, chapter 6, pp. 183–213. John Wiley & Sons, Ltd.

Meyer, J.P., & Zhu, S. (2013). Fair and equitable measurement of student learning in moocs: an introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, *8*, 26–39.

Müller, O., Junglas, I., Brocke, J.V., Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289–302.

Northcutt, C.G., Ho, A.D., Chuang, I.L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers in Education*, *100*(C), 71–80.

O'Neil, C. (2017). Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251). ISSN 0036-8075.

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., Gašević, D. (2016). Generating actionable predictive models of academic performance. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 474–478): ACM.

Pardos, Z.A., Tang, S., Davis, D., Le, C.V. (2017). Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework. In *Proceedings of the fourth (2017) ACM conference on learning @ scale - L@S '17*. ISBN 9781450344500. https://doi.org/10.1145/3051457.3051471.

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., Roll, I. (2017). Identifying productive inquiry in virtual labs using sequence mining. In André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 287–298). Cham: Springer International Publishing.

Qiu, J., Tang, J., Liu, T.X., Gong, J., Zhang, C., Zhang, Q., Xue, Y. (2016). Modeling and predicting learning behavior in moocs. In *Proceedings of the ninth ACM international conference on web search and data mining* (pp. 93–102): ACM.

Reich, J., & Ruipérez-Valiente, J.A. (2019). The MOOC pivot. *Science*, *363*(6423), 130–131.

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. Wiley interdisciplinary reviews: data mining and knowledge discovery, WIREs Data Mining Knowl Discov, 01. https://doi.org/10.1002/widm.1187.

Rosen, Y., Rushkin, I., Ang, A., Federicks, C., Tingley, D., Blink, M.J. (2017). Designing adaptive assessments in MOOCs. In *Proceedings of the fourth (2017) ACM conference on learning @ scale, L@S '17* (pp. 233–236). ISBN 978-1-4503-4450-0.

Ruipérez-Valiente, J.A., Alexandron, G., Chen, Z., Pritchard, D.E. (2016). Using multiple accounts for harvesting solutions in MOOCs. In *Proceedings of the third (2016) ACM conference on learning @ scale - L@S '16* (pp. 63–70).

Ruipérez-Valiente, J.A., Joksimović, S., Kovanović, V., Gašević, D., Muñoz Merino, P.J., Delgado Kloos, C. (2017a). A data-driven method for the detection of close submitters in online learning environments. In *Proceedings of the 26th international conference on world wide web companion* (pp. 361–368).

Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., Alexandron, G., Pritchard, D.E. (2017b). Using machine learning to detect 'multiple-account' cheating and analyze the influence of student and problem features. *IEEE Transactions on Learning Technologies*, *14*(8), 1–11.

Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., Gascón-Pinedo, J.A., Kloos, C.D. (2017c). Scaling to massiveness with ANALYSE: a learning analytics tool for Open edX. *IEEE Transactions on Human-Machine Systems*, *47*(6), 909–914.

Saltelli, A., Chan, K., Scott, E.M., et al. (2000). *Sensitivity analysis* Vol. 1. New York: Wiley.

Seshia, S.A., & Sadigh, D. (2016). Towards verified artificial intelligence. CoRR, arXiv:1606.08514,.

Siemens, G. (2013). Learning analytics: the emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400.

Silver, N. (2012). The signal and the noise: why so many predictions fail–but some don't. Penguin.

U.S. Department of Education, O.iceo.f.E.ducational.T.echnology. (2012). Enhancing teaching and learning through educational data mining and learning analytics: an issue brief.

van der Zee, T., & Reich, J. (2018). Open education science. *AERA Open*, *4*(3), 2332858418787466.

Xing, W., Chen, X., Stein, J., Marcinkowski, M. (2016). Temporal predication of dropouts in moocs Reaching the low hanging fruit through stacking generalization. *Comput. Hum. Behav.*, *58*, 119–129.

Yudelson, M., Fancsali, S., Ritter, S., Berman, S., Nixon, T., Joshi, A. (2014). Better data beats big data. In *Educational data mining 2014*.

Affiliations

Giora Alexandron[1] · Lisa Y. Yoo[2] · José A. Ruipérez-Valiente[2] · Sunbok Lee[3] · David E. Pritchard[2]

Lisa Y. Yoo
lyy@mit.edu

José A. Ruipérez-Valiente
jruipere@mit.edu

Sunbok Lee
sunboklee@outlook.com

David E. Pritchard
dpritch@mit.edu

1 Weizmann Institute of Science, Rehovot, Israel

2 Massachusetts Institute of Technology, Cambridge, MA, USA

3 University of Houston, Houston, TX, USA