International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Defining and measuring completion and assessment biases with respect to English language and development status: not all MOOCs are equal

Sa'ar Karp Gershon[1*], José A. Ruipérez-Valiente[2] and Giora Alexandron[1]

*Correspondence: saar.gershon@weizmann.ac.il
[1] Department of Science Teaching, Weizmann Institute of Science, Rehovot, Israel
Full list of author information is available at the end of the article

Abstract

The emergence of Massive Open Online Courses (MOOCs) broadened the educational landscape by providing free access to quality learning materials for anyone with a device connected to the Internet. However, open access does not guarantee equals opportunities to learn, and research has repetitively reported that learners from affluent countries benefit the most from MOOCs. In this work, we delve into this gap by defining and measuring completion and assessment biases with respect to learners' language and development status. We do so by performing a large-scale analysis across 158 MITx MOOC runs from 120 different courses offered on edX between 2013 and 2018, with 2.8 million enrollments. We see that learners from developing countries are less likely to complete MOOCs successfully, but we do not find evidence regarding a negative effect of not being English-native. Our findings point out that not only the specific population of learners is responsible for this bias, but also that the course itself has a similar impact. Independent of and less frequent than completion bias, we found assessment bias, that is when the mean ability gained by learners from developing countries is lower than that of learners from developed countries. The ability is inferred from the responses of the learners to the course-assessment using item response theory (IRT). Finally, we applied differential item functioning (DIF) methods with the objective of detecting items that might be causing the assessment bias, obtaining weak, yet positive results with respect to the magnitude of the bias reduction. Our results provide statistical evidence on the role that course design might have on these biases, with a call for action so that the future generation of MOOCs focus on strengthening their inclusive design approaches.

Keywords: Bias, MOOC, Human Development Index, Language, Completion, Ability, Item response theory, Odds ratio, Differential item functioning, Inclusive design

Introduction

Massive Open Online Courses provide access to quality learning materials for everyone with an internet connection (Dillahunt & Wang, 2014; Pappano, 2012). Indeed, MOOCs attract a heterogeneous learner population with respect to personal characteristics such as age, level of education, geographic location, and ability (Chuang, 2017; Seaton et al., 2014; Türkay et al., 2017). As an illustrative example, the edX Introductory Physics MOOC 8.MReVx, that is one of the courses studied in this paper, attracted learners from 169 countries, with educational level from secondary or less to advanced degrees, from 14 to 75 years old, and with varying foreknowledge (Chen et al., 2016). As such, MOOCs may be an important vehicle for achieving the United Nations 2030 Agenda for Sustainable Development: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations D.o.E, Affairs S, 2020).

However, open access does not guarantee equal opportunities to learn, and research suggests that learners coming from affluent countries are still the ones who benefit the most from MOOCs (Reich & Ruipérez-Valiente, 2019). MOOCs are dominated by western universities (Adams et al., 2019), and most of the courses are produced in English (Central C, 2021). Proficiency in English and latent cultural and social issues, are embedded as part of the course production and its assessment design. These may affect the ability of global learners to make the most of the new educational opportunities that MOOCs provide (Kizilcec et al., 2017; Türkay et al., 2017). In addition, as MOOC providers put more and more content behind paywalls (Shah, 2017; McKenzie, 2021), MOOCs become less accessible to financially weak learners hence helping to perpetuate the educational gap produced by socioeconomic status (SES). Other demographic factors such as gender are inconclusive with respect to learner success in MOOCs (Rabin et al., 2019), and we expect them to be unrelated to the students ability to learn.

The National Council of Measurement in Education (NCME) defines predictive bias as "The systematic under- or over-prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to the criterion performance." (The National Council of Measurement in Education, 2021). While the NCME's definition is contextualized within standardized summative assessment, we find it also relevant to assessment in MOOCs. In many successful MOOCs (and specifically, the MITx MOOCs that we study), formative assessment also serves for summative purposes. This pedagogy provides learners with resources and incentives for active learning (Alexandron et al., 2020), which is found to be highly effective in MOOCs (Koedinger et al., 2015; Colvin et al., 2014). So, while bias in summative assessment might indicate that students learned less, bias in formative assessment may indicate that the opportunities for learning that these learners received were less adequate to their needs.

Following this, we adopt the term 'bias' in the context of MOOC assessment, and use it to define two measures: *Completion bias* is defined as the reduced likelihood of a subgroup of learners defined by a certain characteristic to complete the MOOC successfully, according to the course definition (typically, achieving a passing grade). Similarly, *assessment bias* is defined as the under-estimation of the ability for learners defined by this characteristic. Specifically, we focus on two characteristics: language, being a native-English-language-speaker (or not) and Human Development Index (HDI), being from a developed or developing country. Our high-level goal is to study whether these characteristics are associated with completion and assessment bias in MOOCs. This goal is formalized into the following research questions (RQs):

RQ1: Do we observe a completion bias with respect to language?

RQ2: Do we observe a completion bias with respect to HDI?
RQ3: Does the course matter with respect to completion bias and HDI?
RQ4: Is there an association between assessment bias and completion bias with respect to HDI?
RQ5: Can differential item functioning (DIF) techniques reduce assessment bias with respect to HDI?

The chain of logic is as follows: First, we measure whether completion bias with respect to language or HDI actually exists. Secondly, we explore whether it is course-related, suggesting that it may be a matter of course design. Thirdly, we aim to analyze the relationship between completion and assessment biases, potentially indicating if the completion bias may be partially connected to increased cognitive difficulties and reduced learning opportunities for certain groups of learners. And lastly, whether we can reduce the assessment bias using item-based analytical methods aimed at identifying problematic items that should be removed or fixed.

To address these questions, we analyze data collected from ~ 150 MITx MOOC runs offered through edX during 2013–2018, with more than two million enrollments and 78,000 certificates earned. Our quantitative methodology draws on learning analytics and psychometrics, and within it, item response theory (IRT) (Meyer & Zhu, 2013) and DIF (Martinková et al., 2017).

The contribution of this paper is threefold: First, it provides large-scale evidence of completion and assessment biases in MOOCs, and sheds some light on the relation between language, HDI, and learning in MOOCs. Secondly, it demonstrates, for the first time, that HDI effects in MOOCs are indeed course-related, and thus may be mitigated with conscious design. Lastly, it suggests a robust psychometrics-based learning analytics methodology, first applied to MOOCs, to address assessment bias and reduce its mean. This methodology can be used to provide MOOC designers with actionable analytics that can assist in reducing both biases in MOOCs.

Literature review

Heterogeneity in MOOCs

Massive Open Online Courses attract learners with unprecedented diversity with respect to factors such as demographic, age, level of education, language, motivation, goals, etc (Chuang, 2017; Seaton et al., 2014; Rabin et al., 2019; Alexandron et al., 2017). Various studies reported the relation between such factors and learning outcomes (Dillahunt & Wang, 2014; Deboer et al., 2013; Morris et al., 2015; Joksimović et al., 2018). These relations are in the focus of MOOC research from its early days: Deboer et al. (2013) explored the first MOOC developed on edX reporting no relation between achievement and age or gender, and only a marginal relation between achievement and level of education. Dillahunt and Wang (2014) incorporated explicitly economic consequences of said related demographics into the analysis of six MOOCs, provided by Coursera. They compared target learners who identify as having financial difficulties, discovering that the target learners are underrepresented in MOOC enrollments, have lower grades, yet more certificates of distinction. Morris et al. (2015) analyzed five MOOCs offered on FutureLearn and found significant association of course completion, age, prior online experience, level of education and employment status. However, due to many anomalies found in the data, these associations were portrayed as simplistic.

This demonstrated diversity in the learner population is the result of MOOCs being a "global classroom" and a movement that raises the flag of democratizing education by providing access to quality learning materials for everyone with a device connected to the internet (Türkay et al., 2017). With respect to this promise, several studies argued that MOOCs actually serve, and are controlled by, the wealthy and educated (Kizilcec et al., 2017; Reich & Ruipérez-Valiente 2019). Other studies were less pessimistic and reported that in some cases learners from developing countries (Zhenghao et al., 2015) or ones who face financial constraints (Dillahunt & Wang, 2014) may actually have higher benefits from MOOCs. Therefore, different studies have brought contradictory views with respect to this story, and our results will further complement this open question.

Language, development status and learning in MOOCs

Achievement gaps may be due to language barriers (Kizilcec et al., 2017), hence language is an important issue. The non-native English speakers are also referred to as English language learners (ELLs), emphasizing that English was learned at some later stage in life and is not a native tongue. Interestingly, even though Joksimović et al. (2018) mentioned that "language represents a primary means of communication in computer-mediated interactions" (p. 68), this systematic literature review makes no reference to language proficiency as a factor that affects student learning in MOOCs. Indeed, Cho and Byun (2017) argued that there is very little evidence on how English as a second language (ESL) participants study in MOOCs. On that, Turkay et al. (2017) provided large-scale evidence of the reduced odds of ELLs to obtain a certificate, and Duru et al. (2019) provided small-scale evidence on the reduced completion rate among ELLs. Another factor that is known to be associated with low learning achievements in MOOCs (and in other contexts) is low SES (Kizilcec et al., 2017). However, the influence of the complex relationship between learning, development and language is still unfolding in MOOC research.

Instructional design that supports the needs of ELLs while learning in a MOOC is required. Resource-oriented approaches that focus on making videos—the primary learning resource in MOOCs (Seaton et al., 2014)—more adapted to ELLs, were presented in Uchidiuno et al. (2018a) and Zee et al. (2017). However, this issue was not explored with relation to assessment, which plays a key role in the learning process of MOOC students (Koedinger et al., 2015; Alexandron et al., 2020).

Similarly to previous studies (Türkay et al., 2017; Uchidiuno et al., 2018b), we refer to language proficiency of an individual as a binary variable with the values Native and Non-native English speaker. Development is measured using the country level Human Development Index (HDI) (Nations U, 2021), as a (very) rough estimate for personal SES. Using HDI is a common practice in MOOC research (Kizilcec et al., 2017; Reich & Ruipérez-Valiente, 2019).

To summarize, not enough is known about language and development effects on MOOCs learning achievements and gaps, and although there is evidence on the effect on certification, there seems to be a gap with respect to the interplay within these two factors and its relation to assessment, as well as evidence that the effect actually differs among courses, hence related to its design. Our work aims to close this gap, by applying IRT and exploring the applicability of DIF methods to MOOCs.

Inferring demographics traits of MOOC learners

The simplest way to infer learners' demographic data is through registration surveys, however the response rate to these surveys in MOOCs is low (Duru et al., 2019). Without direct evidence on enough learners, various approaches are being used to complete the missing demographic information. The common practice for inferring demographics is via Internet Protocol (IP) geolocation mapping (Shavitt and Zilberman, 2011). Language for example is inferred by assuming that learners speak the primary language in the country from which they connect to the platform (Guo and Reninecke, 2014; Seaton et al., 2014). Other practices for learners language inference include using the web browser language preferences (Uchidiuno et al., 2018a), or analyzing the combination of declared locality and keyword search within forum posts (Duru et al., 2019). In a similar fashion, inference of development status for each learner (HDI), is done via IP geolocation mapping and is available at the country level resolution (Nations U, 2021).

Course design and learning achievements

The relation between course design and learning in MOOCs has been mainly studied through between-subjects, within-course analysis, typically in order to optimize learning by identifying causal links between behavior and learning using data mining methods [e.g., Champaign and Cohen (2013), Koedinger et al. (2015), Chen et al. (2016); Renz et al. (2016)]. Within these, several studies highlighted the role of on-going assessment for both measuring (summative) and promoting (formative) learning in MOOCs, as well as for supporting analytics-based design (Alexandron et al., 2020; Koedinger et al., 2015; Reich, 2015). However, there seems to be scarcity of cross-course studies that relate course factors and learning achievements. For example, Joksimović et al. (2018) reviewed only five papers that referred to course characteristics as contextual factors that affect learning, with many more studies looking at learners' variables. To conclude, most relevant research is exploring courses and implicitly including the set of learners in the analysis, not the courses themselves independently of the learner population.

Methodological challenges that between-course studies face might explain this. First, how do we define the instructional design of a MOOC? Such attempts were made by Oh et al. (2020) and Davis et al. (2017), but to-date there is no accepted methodology for encoding MOOC design. Another challenge is the limited access to appropriate data sets. Examples of studies that include multi-course data sets include Oh et al. (2020), Türkay et al. (2017), Miyamoto et al. (2015), Davis et al. (2017) and Ruipérez-Valiente et al. (2020). However, none of these make a cross-course analysis that aims to study the impact of course features on learning outcomes, and indeed, the recent review of Oh et al. (2020) marks this as a valuable research direction. More generally, there is a lack of understanding of the causal factors affecting learning in MOOCs (Joksimović et al., 2018). To address this gap, Joksimović et al. (2018) suggests a unified framework for learning in MOOCs that is based on the model of Reschly & Christenson (2012). This framework seeks to generate associations between student engagement (e.g., motivation, learning goals), contextual factors (e.g., demographic factors and attributes of the course and platform), and learning outcomes (e.g. immediate such as assignments grades, course-level such as course completion). We refer to this framework and demonstrate, to our knowledge for the first time, the relation between course as a whole and learning achievements, using a large-scale cross-course analysis. A research that is closely related to ours is Uchidiuno et al. (2018a), who focused on how ELLs interacted with videos. We focus on interactive assessment and completion rates, and study whether there is evidence that the biases related to these variables are course-related.

Materials and methods

Empirical setup and research population

The empirical setup of our research are 158 MITx MOOC runs of 120 different courses offered on edX between 2013 and 2018, with 2.8 million enrollments and 78,562 certificates awarded. For *RQ1–RQ3*, our research population is composed of the learners who viewed at least once the course materials, denoted as viewers ($N = 1{,}586{,}305$) and the success criterion is course completion, as is defined per MOOC based on its grading policy. For *RQ4*, which conducts a within course analysis, we focus on certified learners within six MOOCs described in Table 1, and then use one of them for *RQ5*; the reasons for the choice of courses are explained in the Results. The unit of analysis for *RQ4–5* is the learner, and is different from that of *RQ1–3*, which is the MOOC.

In terms of content and pedagogy, except for CTL.CFx, the courses followed an active learning pedagogy, with graded formative assessment that focuses on conceptual understanding and is spread along the course. A typical course unit contains three sections: instructional e-text and/or embedded video pages, homework, and a quiz. We expect a considerable variability in the level of English that different problems, and the instructional materials that support them, require. Typically, the problem sets of the instructor-led courses have deadlines, and offer multiple attempts and immediate feedback.

Raw data and processing

The data for this research is based on edX research data exports, which include: (i) data containing user information such as self-reported demographics, course enrollments, etc.; (ii) clickstream data that contain a complete history of user interactions with the courseware; (iii) course metadata files that describe the course elements and the relations between them. These are analyzed using a pipeline combining computations done by *edx2bigquery* (Lopez et al., 2017) which is an open-source package for analyzing edX MOOC data, and a collection of self-developed data-mining scripts.

Inferring learner language and HDI

Each of the methods for inferring these attributes and are mentioned in the literature review has its pros and cons, and we are not familiar with any comprehensive study demonstrating that one of them is superior to the others in the context of MOOCs. Thus, we use the common, IP-based geolocation practice. First, the modal student IP address is used to infer the country to which the student is assigned. We refer to native speakers as ones whose modal IP is assigned to one of the following countries: United States of America, Australia, Canada, Ireland, New Zealand, United Kingdom, Trinidad,

Table 1 RQ4 and RQ5 courses characteristics

	ESD.SCM1x	**8.MReVx**	**7.00x.2**	**JPAL101SPAx**	**CTL.CFx**	**3.MatSelx**
Course title	Supply chain and logistics fundamentals	Mechanics ReView	Introduction to biology—the secret of life	Evaluación de Impacto de Programas Sociales	Supply chain comprehensive exam	Structural materials: selection and economics
Duration (weeks)	13	11 (+ 2 optional)	16	6	1	3
Function	Part of a full online program	Introductory course	Introductory course	NA	Proctored exam for the MITx MicroMasters credential in supply chain management	Engineering introductory course
Instruction language	English	English	English	Spanish	English	English
Instruction model	Instruction led	Instruction led	Instruction led	Self paced	Instruction led	Self paced
N viewed	25,891	15,960	15,712	6041	765	884
n certified	2184	502	940	441	397	884
Gender (%) Male/female/other or missing	70/22/8	72/15/13	51/39/10	44/46/10	74/18/7	54/34/11
Education (%) Secondary/college/advanced	13/45/34	32/32/21	31/31/28	7/46/32	5/42/46	25/24/40
Top three participating countries	US (21%) India (13%) Brazil (4%)	US (25%) India (17%) UK (3%)	US (34%) India (8%) UK (4%)	Peru (15%) Mexico (15%) Chile (15%)	US (25%) India (7%) UK (4%)	US (36%) India (7%) UK (4%)
Total number of countries	201	180	189	87	107	74

and Tobago (The University of Shefield, 2021). Regarding development, the HDI grades countries from developing to developed on the interval [0, 1]. The range [0, 0.7] is defined as developing and describes the majority population in the world; while developed or the minority world is defined as being in the range [0.7, 1], as in Kizilcec et al. (2017). HDI is then used to impute the learners personal SES as either developing or developed. It is important to note that all of the English-native countries are high HDI countries as well, meaning there is not any English-native country with low HDI status.

Measuring academic performance

There are various approaches for measuring learner performance in MOOCs, each raising different pedagogic and methodological issues (Joksimović et al., 2018; Bergner et al., 2015; Alexandron et al., 2020). We use two measures: course completion, and ability estimate that is based on learner responses to the course assessment. Course completion is based on accumulated course grade, which is the predominant proxy for course outcome (Joksimović et al., 2018). Completion is a 'broad brush' binary measure indicating that the learner achieved a passing grade in the course.

Learner ability is estimated using IRT, which is a latent trait model that is based on the idea that the probability of a learner to answer an item correctly is a function of the learner ability and the item difficulty parameters. A main advantage of IRT is that it enables a comparison between learners who attempted different subsets of items (De Ayala and Santiago 2017). We follow the common modeling approach taken in these studies, and use the standard two-parameter logistic model (2PL) item response function (Birnbaum 1968):

$$P_i(s) = \frac{e^{a_i(s-d_i)}}{1 + e^{a_i(s-d_i)}}$$

where d_i and a_i are the difficulty and discrimination parameters of item i, accordingly, and s is the latent learner ability. $P_i(s)$ is thus the probability that a learner with ability s will answer item i correctly.

IRT and MOOC data

Applying IRT to MOOCs raises several methodological challenges, as some of the underlying assumptions may not hold (e.g., a unidimensional, fixed trait). First, IRT assumes that the latent trait is fixed during the activity that generated the responses—an assumption that holds in the context of summative assessment, but not in the context of formative assessment that is spread throughout a MOOC. In this context, the student ability is actually expected to grow (by learning). Nevertheless, the applicability of IRT to MOOC data, and its advantages over classical test theory for measuring learner ability, were demonstrated in several studies (Colvin et al., 2014; Alexandron et al., 2019; Champaign and Cohen, 2013).

Another delicate issue that should be dealt with is modeling multiple attempts, enabled on many of the items in the courses that we study. A common approach, which was also applied to MOOCs, is translating multiple attempts to partial grading models (more attempts equal lower grade), for which there is a plethora of IRT models (Bergner et al., 2015). However, we preferred the more common dichotomous model that is based on correct-on-first-attempt (Alexandron et al., 2016; Champaign and Cohen, 2013), also because multiple learner attempts were found to be highly correlated with cheating (Ruiperez-Valiente et al., 2016).

Another issue is that MOOC data sets typically contain a considerable amount of missing values, as some of the items are randomized, learners may skip items, etc. We removed items that were attempted by less than 10% of the learners, and addressed missing values as Missing-At-Random. In addition, some items may provide very little information on learner ability (e.g., extremely easy or difficult items). We removed such items.

Differential item functioning

DIF refers to a situation in which assessment items function differently across groups (typically demographic) of learners with the same ability level. It is considered a design flaw that may lead to unfair assessment and inaccurate decision making (Martinková et al., 2017). Since learners ability is a measure derived from all items simultaneously, the existence of DIF items may lead to underestimating the ability of some groups of learners. A typical example is a math exam in which some of the items require a considerable reading ability (Haag et al., 2013).

There are several methods for identifying DIF among test items. These can be classified according to several dimensions (Magis et al., 2010). With respect to the methodological dimension, there are two approaches—parametric (IRT based), and non-parametric based. The former rely on the estimation of an IRT model, and the latter is typically based on non-parametric statistical methods for categorical data.

We use non-IRT based methods, as they require less assumptions on the underlying model. Specifically, we use Mantel–Haenszel (MH) (1959) which is a common approach for detecting DIF. It is a Chi-squared contingency table based approach that examines differences between the subgroups on all items of the test, item by item (Marascuilo and Slaughter, 1981). MH divides the ability (operationalized as total score) continuum into K intervals, classifies the examinees into these intervals based on their performance on all the items except for the one being evaluated, and then computes the contingency table per interval, while the MH test itself is for all the tables simultaneously.

We also use logistic regression (LR) which builds a separate LR model per item. The three independent variables are: (a) group membership, (b) total score parameter, and (c) interaction term between the two. The dependent variable is the probability of getting a correct response on the item. Roughly speaking, if the coefficients of the group or interaction terms are significantly different than zero, it is an indication of DIF (non-uniform DIF, in the case of the interaction term). The IRT and DIF analyses were conducted using the R packages *ltm* (Rizopoulos, 2006) and *difR* (Magis et al., 2010).

Analysis and results

RQ1: Do we observe a completion bias with respect to language?

To analyze this, we compute the ratio of two odds: the odds of native English speakers to complete the course successfully, and the odds of non-natives English speakers to do so (the odds are the ratio between the amount of viewers who completed the course and the amount of ones who did not). Hereafter we refer to this quantity as odds ratio. To control for HDI, we focus only on learners from high-HDI countries, as all native English speaking countries we identify as such are characterized by high HDI score (Guo and Reinecke, 2014). See "Inferring learner language and HDI" section for the exact list of countries. Our analysis is based on examining the distribution of the odds ratio in the course runs, as follows. Per course C, we compute a 2×2 contingency table T. The rows of T are defined as the language variable with values $\{native, non-native\}$, and the columns of T are defined as the successful-completion variable in $\{TRUE, FALSE\}$. Concretely, $T[1,\ 1]$—the upper left cell—contains the number of learners who are English-natives and have completed the course, and $T[1,\ 2]$—upper, right—contains the number of English-natives who did not complete the course. Similarly, $T[2,\ 1]$ and $T[2,\ 2]$ contain the amount of non-native speakers who did and did-not, respectively, complete the course. The value of $T[i, j]$ is denoted as n_{ij}. The odds of native speakers to complete C are $\frac{n_{11}}{n_{12}}$, and the odds of non-natives to complete C are $\frac{n_{21}}{n_{22}}$. Using these, the odds-ratio (Bland and Altman, 2000) for C is:

$$\frac{n_{11}}{n_{12}} / \frac{n_{21}}{n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Below, Table 2 is an example of a contingency-table for the “Structural materials: selection and economics” course focusing on course viewers from developed countries. The odds-ratio for this course is 1.14, meaning that being a native speaker increases one’s odds to complete the course by 14%. Following the standard methodology for analyzing odds ratio, hypothesis testing is done on the log of the odds ratio, which is normally distributed under the null hypothesis.

The distribution of the log odds ratio over the 159 course runs is presented in Fig. 1 in a bold line, and a reference distribution indicating no effect is in dotted line. To test whether language and completion are independent (given high HDI), we test whether the sample mean of the log odds ratio is different than zero (and respectively, the odds ratio is compared to one, meaning equal odds). The odds ratio is distributed with mean $M = 1.076$, which is insignificantly different from 1 with $p = 0.464$. Thus, we conclude that the likelihood of completing the course is similar for both groups, hence completion is independent of the language.

RQ2: Do we observe a completion bias with respect to HDI?

Next, we look at the HDI variable and ask whether it is independent of the completion variable. To control for the language variable, we include only the non-native English speakers in this analysis. We follow the same procedure as in *RQ1*, with HDI replacing the language variable having possible values “high” and “low.” Thus, per course C, the contingency table T is defined as follows: $T[1,\ 1]$ contains the number of learners who are high HDI and have completed the course, and $T[1,\ 2]$ contains the number of learners who are high HDI, and did not complete the course. Similarly, $T[2,\ 1]$ and $T[2,\ 2]$ contain the number of low HDI learners who did and did-not complete the course, respectively. This is demonstrated in Table 3, for the course Structural materials: selection and economics.

Table 2 Contingency table of “Structural materials: selection and economics”, third trimester, 2017, successful completion by language

	Completed	Did not complete	
Native speaker	61	572	633
Non-native speaker	117	1252	1369
	178	1824	2002

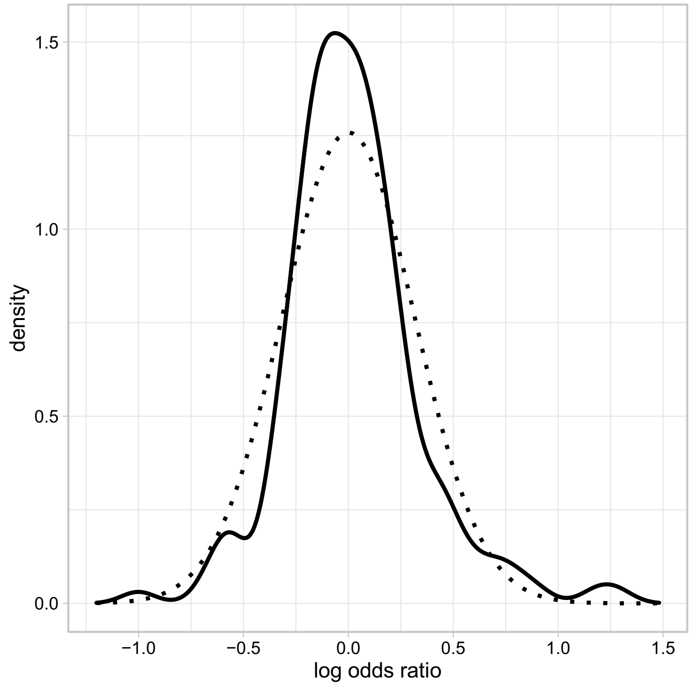

Fig. 1 Comparison of the no-effect and the sample log odds ratio distributions by language (full = sample, dotted = no-effect)

Table 3 Contingency table of "Structural materials: selection and economics", third trimester, 2017, successful completion by HDI

	Completed	**Did not complete**	
High HDI	117	1252	1369
Low HDI	64	763	827
	181	2015	2196

The odds ratio for this course is 1.11, meaning that being from a high HDI country increases one's odds to complete the course by 11%. To test whether HDI and completion are independent, we again analyze whether the mean of the log odds ratio is statistically different than zero. Due to statistical considerations regarding minimal counts per cell ($n_{i,j}$), the total number of analyzed courses is slightly different than that of RQ1, altogether 137 course runs. The distribution of the log odds ratio among the 137 course runs is presented in Fig. 2, together with the hypothetical no-effect density (dotted line). The mean odds ratio $M = 1.83$, is significantly different from 1 with $p < .001$. In other words, controlling for language by focusing on learners from non English-native countries, the completion odds of a learner from high HDI countries are 1.83 times larger than the odds of a learner from low HDI countries.

To conclude the findings of RQ1 and RQ2 we indicate that being English-native and completing the course are independent of each other within learners from developed countries, while there is an association between HDI and completing the course within the group of non English-native learners. Hence, for the next question we focused our

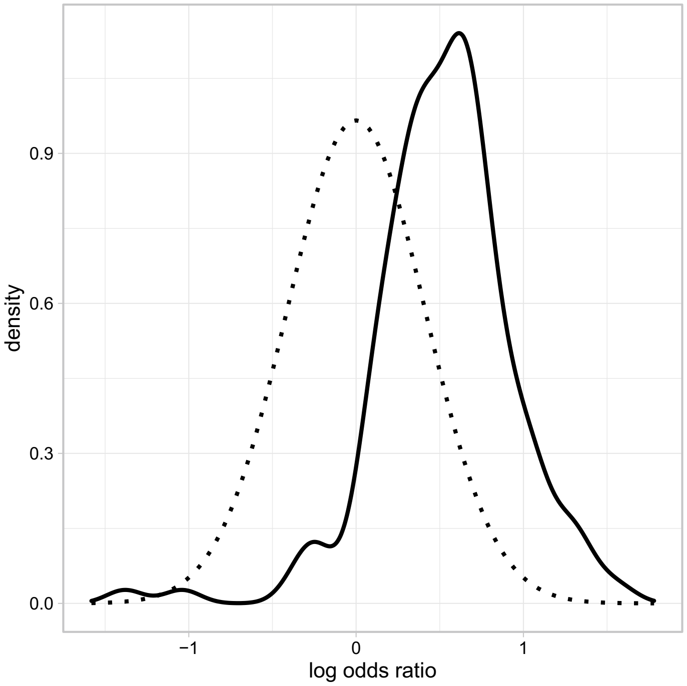

Fig. 2 Comparison of the no-effect and the sample log odds ratio distributions by development and controlled for language (full = sample, dotted = no-effect)

attention on non English-native learners, and proceed with examining the bias defined by HDI and measured by the odds ratio as in RQ2.

RQ3: Does the course matter with respect to completion bias and HDI?

The rationale that drives our analysis is that if the "course matters" in terms of HDI-related completion bias (demonstrated in RQ2), the odds ratio of runs of the same course (for courses offered several times) would be more similar than runs of different courses. To control for language, we include only the non-native English speakers. Our assumption is that between runs of the same course, only the population changes, while between different courses, both the population and the course change. In any case, we assume that the learners of each course-run are a random sample from a general population with respect to the measured variable, course completion. To analyze whether course and HDI completion bias are independent, we perform a permutation test. We compute the pairwise distance (absolute distance) d between the log odds ratio of each pair of runs i, j, denoted $d(i, j)$. We then define two groups: $\text{group A} = \{d(i,j)|i \text{ and } j \text{ are different runs of the same course}\}$, and $\text{group B} = \{d(i,j)|i \text{ and } j \text{ are runs of different courses}\}$.

The distribution of $d(i, j)$ for groups A and B is presented in Fig. 3 (A = True, B = False), where the bold lines represent the medians of each distribution. Looking at the within-course pairwise distances, the mean of log(odds ratio)s absolute difference of completion is $M = 0.16$ based on $n = 124$ pairs and $Mdn = 0.05$. For pairs of different courses the mean is $M = 0.3$ based on $n = 1992$ pairs and $Mdn = 0.26$. The difference between the means is significant with $p < .001$ in a Welch Two Sample t-test with

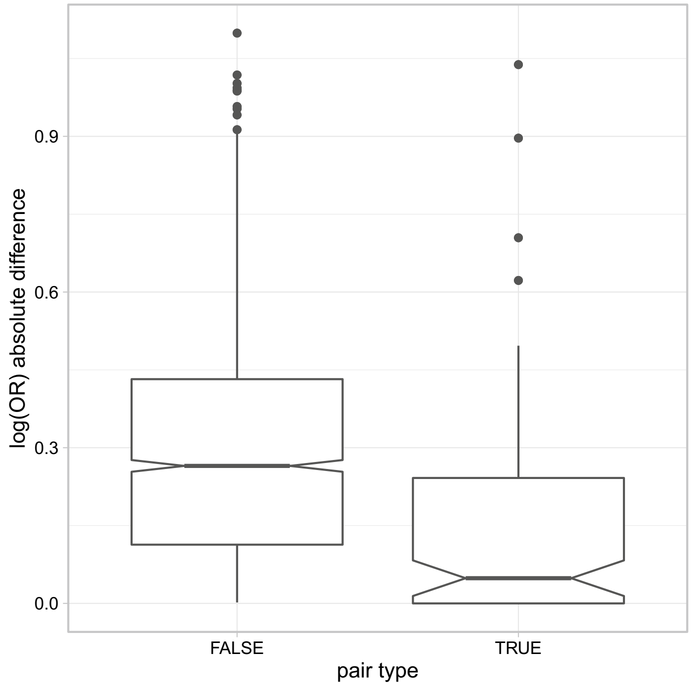

Fig. 3 The distribution of absolute differences between log odds ratios of all pairs by pair type

$t(137.55) = 6.01$. The notches of the boxes in the box-plot in Fig. 3 indicate a strong evidence that the medians of these distributions differ significantly, as well. Since the changes are absolute we can infer that course related factors contribute for a mean change of $0.3 - 0.16 = 0.14$. Thus, we conclude that "the course matters" with respect to completion bias.

RQ4: Is there an association between assessment bias and completion bias with respect to HDI?

As the course explains almost half of the value of the odds ratio (47%), we turned our attention to examine which components of the course may contribute to the completion bias. As the assessment determines learners' grade and thus determines the completion status, we examined whether we can find similar signs of bias against low HDI learners in the assessment, which may explain the bias in the completion rates. To do so, we analyze six courses—the three courses with the largest number of certified learners among the courses that include completion bias (SCM, 7.00, 8MReV), and the three courses with the largest number of certified learners among the courses that do not include completion bias (JPAL101SPAx, CTL.CFx, 3.MatSelx). The characteristics of these courses are described in Table 1. As an anecdote, the course JPAL101SPAx is using Spanish as the instruction language, which can be an explanation for the lack of completion bias. We note that even though completion bias exists on the level of the entire course portfolio, individual courses differ on this measure, with some containing no bias at all.

To measure the assessment bias (defined as the under-estimation of the ability for learners from developing countries; see "Introduction", paragraph above the RQ's), we compare the ability estimate, computed using IRT, of the two groups. The distribution

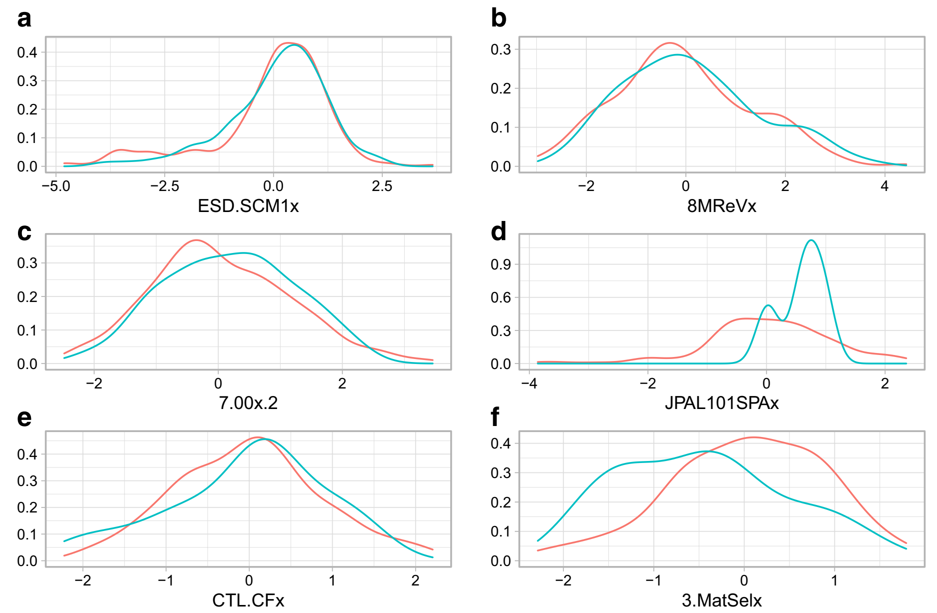

Fig. 4 Estimated IRT ability parameters by HDI (blue = developing, red = developed)

Table 4 RQ4 and RQ5 (bolded half column titled 3.MatSelx) courses completion and assessment bias characteristics

	ESD.SCM1x	**8.MReVx**	**7.00x.2**
Developed mean	− 0.048	− 0.041	0.04
Developing mean	0.089	0.095	0.183
t (df)	0.98 (206.73)	0.79 (191.67)	1.12 (132.1)
p	.331	.434	.266
OR (p)	1.684 (< .001)	1.442 (.004)	1.248 (.029)
n (learners/items)	531/468	501/434	940/853
	JPAL101SPAx	**CTL.CFx**	**3.MatSelx**
Developed mean	0.053	− 0.001	**0.033**
Developing mean	0.582	− 0.043	**− 0.461**
t (df)	2.375 (4.18)	0.247 (55.43)	**2.388 (37.38)**
p	.074	.806	**.022**
OR(p)	1.207 (.248)	1.097 (.622)	**1.114 (.505)**
n (learners/items)	160/300	348/65	**152/41**

of the ability estimates is presented in Fig. 4. Among the six courses, five do not contain assessment bias. The sixth course and the only one that does contain assessment bias, is 3.MatSelx, Facet f, Fig. 4. However this course does not demonstrate completion bias. The characteristics and inference results of the bias analysis are summarized in Table 4.

Regarding 3.MatSelx which exhibited assessment bias, the mean ability of the learners from developed countries is M = 0.03, whereas M = − 0.46 for the developing countries. This difference is statistically significant with $t(37.38) = 2.39, p = 0.022$, meaning there is a significant assessment bias. Thus, the assessment bias in this course-run can be calculated as 0.03 − (−0.46) = 0.49 measured in standard deviations. To conclude the results of RQ4, we see no evidence that completion bias and assessment bias are systematically associated.

RQ5: Can DIF techniques reduce assessment bias with respect to HDI?

Last, and following the findings indicating that assessment bias may exist in some courses, we turn our attention to studying methods that may reduce the assessment bias. We focus on methods that aim to identify items on which there is a statistically significant difference between different sub-groups, collectively known as DIF, as described in "Materials and methods" section. We ran these methods on 3.MatSelx, in which we observed a statistically significant assessment bias against low HDI learners. The final IRT model of this course is based on 152 learners out of the 241 learners who achieved a passing grade (completers).

Table 5 presents the number of problematic 'biased' items that were identified by each method, and the effect of removing these items on the mean value of the ability estimate for each group.

The number of items in the full IRT model that is used to demonstrate the assessment bias contains 36 items. The removed items using each method turned up to be not unique: MH yields a single item, and the Logistic method yields three items, among them the one identified by MH.

Based on our DIF results, the logistic-based analysis is the most effective in discovering problematic items, and in narrowing the gap between the groups. However, even after removing the items discovered by these methods, the difference between the means of ability estimate in the groups (assessment bias) is still statistically significant (p = .035). The assessment bias is still present after the DIF treatment.

Summary of results

To conclude this section, we see that:

- The odds of learners from low HDI countries to complete MOOCs successfully are 55% of those of learners from high HDI countries.
- With respect to language, we see no evidence that learners whose native language is not English have lower odds to complete MOOCs successfully when controlling for development status.

Table 5 DIF methods on the course 3.MatSelx

Method	Original model	MH	Logistic
Number of items in Model	36	35	33
Number of removed items (%)	–	1 (3%)	3 (8%)
Mean ability for developed	0.03	0.02	0.01
Mean ability for developing	− 0.46	− 0.48	− 0.44
Difference between means (%)	0.49	0.5 (102%)	0.45 (92%)

- The course itself is also responsible for the effect on learners from low HDI countries, explaining 47% of the value of the odds ratio.
- The completion bias is not associated with assessment bias among the successful course-completers and in 83% of the MOOCs we checked we see no evidence that learners from low HDI countries are disadvantaged. Assessment bias may be found in some courses, specifically in 17% of the courses we looked at.
- Applying DIF methods for addressing the assessment bias found in a course yields weak results with respect to the magnitude of reduction.

Discussion, limitations and conclusions

MOOCs promise access to high quality learning materials for everyone with an internet connection. However, our findings demonstrate that the odds of non-native English speakers from low HDI countries to complete a MOOC are nearly half of the ones of non-native English learners from high HDI countries (RQ2); we refer to this as completion bias with respect to HDI. The language variable, when controlling for development, was found to be independent of MOOC completion, which contradicts (Türkay et al., 2017), which reported that the odds of ELLs to obtain a certificate are $\sim 60\%, p < .001$ of those of native speakers.

The contradicting results can be explained by many factors: (a) The measured outcomes—completion and certification—are strongly correlated, but not identical; (b) The data sets are different, although both are large MOOCs datasets (~ 150 MITx vs. ~ 100 HarvardX MOOCs); (c) The model and control variables are different; (d) The results of Türkay et al. (2017) are reported on the pooled data of all learners in the MOOCs whereas our results are computed per MOOC, therefore Simpson's paradox (Blyth, 1972) can explain the different results; (e) We used the data of all the learners and computed the language/HDI variables based on IP, while the results of Türkay et al. (2017) are based on the edX survey, which are typically answered only by a small subset of the MOOC learners, and are thus subject to self-selection bias. These different results call for additional research on the role that language may play in MOOCs, as this can inform more equitable course production across providers and improve the learning experience in MOOCs for all learners.

Aside from the relation between completion, language and development status, another question arises regarding the factors affecting the completion bias. What is the effect on the completion bias of within-course features such as assessment items, and of across-course-run features such as learner population. The analysis of RQ3 underlines that course features are an important factor affecting the completion bias, indicating that *the course design may strongly affect the odds of non-native speakers from low HDI countries to complete MOOCs*. This first (to the best of our knowledge) statistical evidence on the role of course design on course completion gives hope that some of the bias may be treated with more inclusive design. Similarly, previous large-scale interventions have shown that these might work in some courses but not in others, also suggesting the importance of course design and learner population (Kizilcec & Kambhampaty, 2020). The future generation of MOOCs should strengthen their inclusive design approaches (Iniesto and Rodrigo, 2018).

In light of these findings on the effect of course design on completion odds of learners from developing countries, we focused our attention on the assessment. The results of RQ4 showed that the relation between ability estimate and course completion might be far more intricate than we might have anticipated. Out of six courses that we examined, only one exhibited assessment bias, even though this course was without significant evidence of completion bias in it. In four of six (67%) courses the difference in ability by HDI is actually opposite from expected, with low HDI learners achieving *greater* mean ability than high HDI ones, although all differences are statistically insignificant. Our IRT methodology resulted in a unique group of learners that represent the core learning group in the course-run hence allowing to shed some light on the relation between ability estimate and completion rate. Assessment serves both learning and measurement purposes, and may be the single most important factor affecting learning in MOOCs (Koedinger et al., 2015; Colvin et al., 2014; Alexandron et al., 2020). If learners from developing countries face increased difficulty during learning, they will be more likely to drop-out, although the ones that “survive”, do not “pay” in a decreased estimate of their ability, usually, and sometimes excel over the learners from developed countries. This excellence might be the result of additional motivation and dedication of learners from developing countries although this remains an open question.

In RQ5 we explored means to treat the observed assessment bias, by identifying problematic items that are biased with respect to HDI. While our IRT-oriented, DIF-based methods did identify up to 8% of the items as problematic, removing these items had only a minor effect on the difference between the ability estimate of both groups (Table 5). This may have several explanations. Firstly, there could be a few methodological explanations: The non-parametric DIF methods that we have used require less assumptions than parametric ones, but are of lower power, meaning that only items that are severely biased are detected. In addition, it might be that there is just too much bias and heterogeneity for such methods to work properly—DIF methods look for a contrast between the behavior of a single item and the overall trend. If too many items are biased, this item-level approach would fail. However, we believe that there might be a deeper conceptual issue involved. DIF is a reductionist, bottom-up approach that operates under the assumption that course elements can be treated separately. But (effective) learning design may have a more holistic nature, with intricate relations between the course elements. In such cases, reductionist approaches might be inappropriate, and treating the bias effectively would require top-down or even better, an integrated approaches. Only after such approaches are applied, item-level methods can be applied to further optimize the assessment.

To conclude the discussion we borrow terms and concepts from Joksimović et al. (2018) who offered a framework for learning in MOOCs. Both biases we explored, completion and assessment, are two different complementary evidences of learning, each from a different perspective. Course completion is a conventional course-level academic learning outcome, usually composed of success and participation measures of the course content as is indicated by the existence of a grading formula. Our results indicate that not only is it different for distinct learners based on demographic contexts (development status), it is also associated with the course content, and specifically the assessment of the course. The one size fits all model of MOOC pedagogic design creates evidently different sorts of engagement in the course; explicitly, the assessment bias which is an immediate, academic, and direct learning outcome (evident on learners-groups formed using the IP coded language and development status). The minor reduction in assessment bias contributed to using DIF suggests that although IRT captures important aspects of engagement and learning in MOOCs, course design and pedagogy are of a more holistic nature and are not fully explained by its item-wise approach.

Our study has several limitations: with respect to measurement error, we use the learner's modal IP as a proxy for language and development status. This methodology has obvious limitations such as misclassification, where non-natives (natives) reside in an English (non-English) speaking country or use IP concealing tech (i.e., Virtual Private Network (VPN)). Still, it is a common methodology within online (education) research, which both the literature and our belief indicate as of sufficient accuracy for our purposes. This answers a similar argument that can be said regarding the country level development status which neglect within country variability and instead use a country-wide average. With respect to the data, our study is based on a database of MITx MOOCs. This is *not* a randomly selected set of MOOCs in terms of domain (almost entirely science and technology), course design (e.g., MITx MOOCs may be more problem-solving oriented), and the learner population that these courses may attract.

Conclusion

This paper analyzed the association between language, development status, and learning achievements in MOOCs, using a large data set of 158 MITx MOOC runs from 120 different courses offered on edX between 2013 and 2018, with 2.8 million enrollments. The results yielded that with respect to language, when controlling for development status, native and non-native English speakers have similar odds to complete the course. However with respect to development status, when controlling for language, the results indicate a clear bias against learners from low HDI countries, whose odds to complete the course are only 55% of learners from high HDI countries. Next, we explored the role of the course in explaining the completion bias, and found that the course explains 47% of the value of the odds ratio. However, and quite surprisingly, we did not find an association between said completion bias and the bias in the ability estimate of learners from low HDI countries who did complete the course. Within a course that does include such assessment bias, psychometrics techniques for eliminating bias in assessment were marginally effective.

Our findings significantly strengthen previous results on the negative association between development status and performance in MOOCs, and contradict the results of previous studies that reported that non-native English speakers have lower odds to succeed in MOOCs, and they do so based on a much larger data set. In addition, they provide the first large-scale statistical evidence on the role that course design might have on these biases. These findings both call for action for closing these gaps, and give hope that inclusive design in MOOCs may actually help to address them. On the methodological aspect, this work defines two useful terms—completion and assessment bias, and powerful methodologies to study them, the odds ratio and IRT. In addition, it is the first to examine the applicability of DIF methods for the purpose of increasing the fairness of assessment in MOOCs by advising on the course design.

Abbreviations

DIF: Differential item functioning; ELL: English language learner; ESL: English as a second language; HDI: Human Development Index; IP: Internet Protocol; IRT: Item response theory; LR: Logistic regression; MH: Mantel–Haenszel; MOOC: Massive Open Online Course; NCME: The National Council of Measurement in Education; RQ: Research question; SES: Socioeconomic status; VPN: Virtual Private Network.

Acknowledgements

The authors would like to thank MITx, and especially Jon Daries, for providing access to the data, and to Aviram Berg for his contribution in the data-curation stages of this research.

Authors' contributions

SKG was lead in analysis, investigation, conceptualization, methodology, project administration, software development, visualization, and writing of the first draft; equal contributor to data curation, validation and the editing and reviewing and a supporting role to funding acquisition. JARV was equal contributor to the validation and the editing and reviewing and had a support role to the writing of the first draft, visualization and supervision. GA was lead of resources, supervision and funding acquisition, equal contributor to the data curation, validation, and editing and reviewing, and a support role to the conceptualization, methodology, project administration, software development, visualization, and writing the first draft. All authors read and approved the final manuscript.

Funding

The work of SKG was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center. The Israeli Ministry of Science and Technology supported the research of GA under Project no. 3-15607.

Availability of data and materials

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interest

The authors declare that they have no competing interests.

Author details

[1]Department of Science Teaching, Weizmann Institute of Science, Rehovot, Israel. [2]University of Murcia, Calle Campus Universitario, Murcia, Spain.

Received: 15 March 2021 Accepted: 19 May 2021
Published online: 27 July 2021

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology, 72*(3), 466–485. https://doi.org/10.1111/bmsp.12169.

Alexandron, G., Lee, S., Chen, Z., & Pritchard, D. E. (2016). Detecting cheaters in MOOCs using item response theory and learning analytics. In *CEUR workshop proceedings*, vol. 1618.

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. *Computers and Education, 108,* 96–114. https://doi.org/10.1016/j.compedu.2017.01.015.

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 *EDM 2019—Proceedings of the 12th international conference on educational data mining (Edm)* (pp. 480–483). https://doi.org/10.35542/osf.io/wuqv5.

Alexandron, G., Wiltrout, M. E., Berg, A., & Ruipérez-Valiente, J. A. (2020). Assessment that matters: Balancing reliability and learner-centered pedagogy in MOOC assessment. In *ACM international conference proceeding series* (pp. 512–517). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3375462.3375464.

Bergner, Y., Colvin, K., & Pritchard, D. E. (2015). Estimation of ability from homework items when there are missing and/or multiple attempts. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 118–125).

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ, 320*(7247), 1468.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366.

Central, C. Languages. Retrieved January 25, 2021 from https://www.classcentral.com/languages.

Champaign, J., & Cohen, R. (2013). Ecological content sequencing: From simulated students to an effective user study. *International Journal of Learning Technology, 8*(4), 337. https://doi.org/10.1504/ijlt.2013.059130.

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*. https://doi.org/10.1186/s41039-016-0034-4.

Cho, M.-H., & Byun, M. (2017). International review of research in open and distributed learning IRRODL nonnative English-speaking students' lived learning experiences with MOOCs in a regular college classroom nonnative english-speaking students' lived learning experiences with MOOCs in a regular college classroom. *International Review of Research in Open and Distributed Learning, 18*(5), 173–190. https://doi.org/10.19173/irrodl.v18i5.2892.

Chuang, I. (2017). HarvardX and MITx: Four years of open online courses—Fall 2012-Summer 2016. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2889436.

Colvin, K. F., Champaign, J., Liu, A., Zhou, Q., Fredericks, C., & Pritchard, D. E. (2014). Learning in an introductory physics mooc: All cohorts learn equally, including an on-campus class. *The international review of research in open and distributed learning*, *15*(4).

Davis, D., Jivet, I., Kizilcec, R. F., Chen, G., Hauff, C., & Houben, G. -J. (2017). Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 454–463).

De Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*. https://doi.org/10.1016/j.jsp.2016.01.002.

Deboer, J., Seaton, D. T, & Breslow, L. (2013). Diversity in MOOC students ' backgrounds and behaviors in relationship to performance in 6.002x. In *Proceedings of the sixth learning international networks consortium conference* (pp. 1–10). Retrieved from https://www.researchgate.net/publication/237092327.

Dillahunt, T., & Wang, Z. (2014). Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *International Review of Research in Open and Distance Learning, 15*(5), 177–196. https://doi.org/10.19173/irrodl.v15i5.1841.

Duru, I., Sunar, A. S., White, S., Diri, B., & Dogan, G. (2019). A case study on English as a second language speakers for sustainable MOOC study. *Sustainability (Switzerland), 11*(10), 2808. https://doi.org/10.3390/su11102808.

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Association for computing machinery*, New York, New York, USA. https://doi.org/10.1145/2556325.2566247.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28,* 24–34. https://doi.org/10.1016/j.learninstruc.2013.04.001.

Iniesto, F. & Rodrigo, C. (2018). Yourmooc4all: A MOOCs inclusive design and useful feedback research project. In *2018 learning with MOOCs (LWMOOCS)* (pp. 147–150). https://doi.org/10.1109/LWMOOCS.2018.8534644.

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., et al. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research, 88*(1), 43–86. https://doi.org/10.3102/0034654317740335.

Kizilcec, R. F., & Kambhampaty, A. (2020). Identifying course characteristics associated with sociodemographic variation in enrollments across 159 online courses from 20 institutions. *PLoS ONE, 15*(10), 0239766. https://doi.org/10.1371/journal.pone.0239766.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science, 355*(6322), 251–252. https://doi.org/10.1126/science.aag2063.

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. In *L@S 2015—2nd ACM conference on learning at scale* (pp. 111–120). https://doi.org/10.1145/2724660.2724681

Lopez, G., Cambridge, H., Seaton, D.T., Ang, A., Tingley, D., & Chuang, I. (2017). Google BigQuery for education: Framework for parsing and analyzing edX MOOC Data. In *Proceedings of the fourth (2017) ACM conference on learning @ scale*. ACM, New York, NY, USA. http://dx.doi.org/10.1145/3051457.3053980.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ2 statistics. *Journal of Educational Measurement, 18,* 229–248.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sciences Education, 16*(2), 2. https://doi.org/10.1187/cbe.16-10-0307.

McKenzie, L. Is a shakeout coming for online program management companies? Retrieved January 25, 2021 from https://www.insidehighered.com/digital-learning/article/2018/06/04/shakeout-coming-online-program-management-companies.

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.

Miyamoto, Y. R., Coleman, C., Williams, J. J., Whitehill, J., Nesterko, S., & Reich, J. (2015). Beyond time-on-task: The relationship between spaced study and certification in MOOCs. *Journal of Learning Analytics, 2*(2), 47–69. https://doi.org/10.18608/jla.2015.22.5.

Morris, N. P., Morris, N. P., Hotchkiss, S., & Swinnerton, B. (2015). Can demographic information predict MOOC learner outcomes? Can demographic information predict MOOC learner outcomes? *Proceedings of the European MOOC Stakeholder Summit, 2015*(MAY) (pp. 199–207).

National Council on Measurement in Education. Assessment glossary. Retrieved January 25, 2021 from https://www.ncme.org/resources/glossary.

Nations, U. Human Development Index (HDI) | Human development reports. Retrieved January 29, 2021 from http://hdr.undp.org/en/content/human-development-index-hdi.

Oh, E. G., Chang, Y., & Park, S. W. (2020). Design review of MOOCs: Application of e-learning design principles. *Journal of Computing in Higher Education, 32*(3), 455–475. https://doi.org/10.1007/s12528-019-09243-w.

Pappano, L. (2012). The year of the mooc. *The New York Times, 2*(12), 2012.

Rabin, E., Henderikx, M., Kalman, Y. M, & Kalz, M. (2019). The influence of self-regulation, self-efficacy and motivation as predictors of barriers to satisfaction in moocs. In *European conference on technology enhanced learning* (pp. 631–635). Springer.

Reich, J. (2015). Rebooting MOOC research: Improve assessment, data sharing, and experimental design. *Science, 347*(6217), 34–35. https://doi.org/10.1126/science.1261627.

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science, 363*(6423), 130–131. https://doi.org/10.1126/science.aav7958.

Renz, J., Hoffmann, D., Staubitz, T., & Meinel, C. (2016). Using A/B testing in MOOC environments. In *ACM international conference proceeding series*, vol. 25–29-April-2016 (pp. 304–313). Association for Computing Machinery, New York, New York, USA. https://doi.org/10.1145/2883851.2883876. http://dl.acm.org/citation.cfm?doid=2883851.2883876.

Reschly, A. L., & Christenson, S. L. (2012). Moving from "context matters" to engaged partnerships with families. *Journal of Educational and Psychological Consultation, 22*(1–2), 62–78. https://doi.org/10.1080/10474412.2011.649650.

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Ruiperez-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* (pp. 63–70).

Ruipérez-Valiente, J. A., Jenner, M., Staubitz, T., Li, X., Rohloff, T., Halawa, S., Turro, C., Cheng, Y., Zhang, J., Zabala, I. D., Reich, J., Despujol, I., & Reich, J. (2020). Macro MOOC learning analytics: Exploring trends across global and regional providers. In *ACM international conference proceeding series (December)* (pp. 518–523). https://doi.org/10.35542/osf.io/9ghfc.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM, 57*(4), 58–65. https://doi.org/10.1145/2500876.

Shah, D. MOOCs 2017: A year in review by class central. Retrieved January 25, 2021 from https://www.classcentral.com/moocs-year-in-review-2017.

Shavitt, Y., & Zilberman, N. (2011). A geolocation databases study. *IEEE Journal on Selected Areas in Communications, 29*(10), 2044–2056. https://doi.org/10.1109/JSAC.2011.111214.

The University of Sheffield. List of majority native English speaking countries | International students. Retrieved February 02, 2021 from https://www.sheffield.ac.uk/international/english-speaking-countries.

Türkay, S., Seaton, D., Eidelman, H., Lopez, G., Rosen, Y., & Whitehill, J. (2017). Getting to know English language learners in MOOCs: Their motivations, behaviors and outcomes. In *L@S 2017—Proceedings of the 4th ACM conference on learning at scale* (pp. 209–212). https://doi.org/10.1145/3051457.3053987.

Uchidiuno, J., Koedinger, K., Hammer, J., Yarzebinski, E., & Ogan, A. (2018a). How do English language learners interact with different content types in MOOC videos? *International Journal of Artificial Intelligence in Education, 28*(4), 508–527. https://doi.org/10.1007/s40593-017-0156-x.

Uchidiuno, J. O., Ogan, A., Yarzebinski, E., & Hammer, J. (2018b). Going global: Understanding English language learners' student motivation in English-language MOOCs. *International Journal of Artificial Intelligence in Education, 28*(4), 528–552. https://doi.org/10.1007/s40593-017-0159-7.

United Nations, D.o.E., Affairs, S. Sustainable Development Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. Retrieved May 01, 2020 from https://sdgs.un.org/goals/goal4.

Zee, T. V. D., Admiraal, W., Paas, F., Saab, N., & Giesbers, B. (2017). Effects of subtitles, complexity, and language proficiency on learning from online education videos. *Journal of Media Psychology, 29*(1), 18–30. https://doi.org/10.1027/1864-1105/a000208.

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D, & Emanuel, E. J. (2015). Who's benefiting from MOOCs, and why. Retrieved January 25, 2021 from https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.