Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions

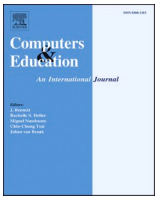
José A Ruipérez-Valiente, Thomas Staubitz, Matt Jenner, Sherif Halawa, Jiayin Zhang, Ignacio Despujol, Jorge Maldonado-Mahauad, German Montoro, Melanie Peffer, Tobias Rohloff, et al.

Computers & Education 180 (2022) 104426

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu



Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions

José A. Ruipérez-Valiente [a,b,*], Thomas Staubitz [c], Matt Jenner [d], Sherif Halawa [e], Jiayin Zhang [f], Ignacio Despujol [g], Jorge Maldonado-Mahauad [h,i], German Montoro [j], Melanie Peffer [k], Tobias Rohloff [c], Jenny Lane [k], Carlos Turro [g], Xitong Li [l], Mar Pérez-Sanagustín [h,m,n], Justin Reich [b]

[a] University of Murcia, Murcia, Spain
[b] Massachusetts Institute of Technology, Cambridge, MA, USA
[c] Hasso Plattner Institute, Potsdam, Germany
[d] FutureLearn, London, UK
[e] Edraak, Amman, Jordan
[f] Tsinghua University, Beijing, China
[g] Universitat Politècnica de València, Valencia, Spain
[h] Pontificia Universidad Católica de Chile, Santiago de Chile, Chile
[i] Universidad de Cuenca, Cuenca, Ecuador
[j] Universidad Autónoma de Madrid, Madrid, Spain
[k] University of Colorado Boulder, Boulder, CO, USA
[l] HEC Paris, Jouy-en-Josas, France
[m] Université Paul Sabatier de Toulouse III, Toulouse, France
[n] Institute de Recherche en Informatique de Toulouse (IRIT), Toulouse, France

ARTICLE INFO

Keywords:
Learning analytics
Educational data mining
Massive open online courses
Large scale analytics
Cultural factors
Equity
Distance learning

ABSTRACT

Massive Open Online Courses (MOOCs) remarkably attracted global media attention, but the spotlight has been concentrated on a handful of English-language providers. While Coursera, edX, Udacity, and FutureLearn received most of the attention and scrutiny, an entirely new ecosystem of local MOOC providers was growing in parallel. This ecosystem is harder to study than the major players: they are spread around the world, have less staff devoted to maintaining research data, and operate in multiple languages with university and corporate regional partners. To better understand how online learning opportunities are expanding through this regional MOOC ecosystem, we created a research partnership among 15 different MOOC providers from nine countries. We gathered data from over eight million learners in six thousand MOOCs, and we conducted a large-scale survey with more than 10 thousand participants. From our analysis, we argue that these regional providers may be better positioned to meet the goals of expanding

* Corresponding author. University of Murcia, Murcia, Spain.
E-mail addresses: jruiperez@um.es (J.A. Ruipérez-Valiente), thomas.staubitz@hpi.de (T. Staubitz), matt.jenner@futurelearn.com (M. Jenner), shalawa@qrf.org (S. Halawa), zhangjy5@sem.tsinghua.edu.cn (J. Zhang), ndespujol@asic.upv.es (I. Despujol), jorge.maldonado@ucuenca.edu.ec (J. Maldonado-Mahauad), german.montoro@uam.es (G. Montoro), Melanie.Peffer@colorado.edu (M. Peffer), Tobias.Rohloff@hpi.de (T. Rohloff), Jenny.Lane@colorado.edu (J. Lane), turro@cc.upv.es (C. Turro), lix@hec.fr (X. Li), mar.perez-sanagustin@irit.fr (M. Pérez-Sanagustín), jreich@mit.edu (J. Reich).

https://doi.org/10.1016/j.compedu.2021.104426
Received 15 April 2021; Received in revised form 23 December 2021; Accepted 27 December 2021
Available online 7 January 2022

access to higher education in their regions than the better-known global providers. To make this claim we highlight three trends: first, regional providers attract a larger local population with more inclusive demographic profiles; second, students predominantly choose their courses based on topical interest, and regional providers do a better job at catering to those needs; and third, many students feel more at ease learning from institutions they already know and have references from. Our work raises the importance of local education in the global MOOC ecosystem, while calling for additional research and conversations across the diversity of MOOC providers.

1. Introduction

During the last decade, Massive Open Online Courses (MOOCs) have settled within the higher education ecosystem creating new opportunities and business models (Reich & Ruipérez-Valiente, 2019; Yuan & Powell, 2013). The initial promise of MOOCs involved reaching the entire world through making free courses from elite universities available online (Dillahunt et al., 2014). However, despite the tremendous opportunities that MOOCs represent for human development (Zhang et al., 2019), previous studies have suggested that the majority of learners attracted by MOOC providers are already educated and come from affluent countries (Reich & Ruipérez-Valiente, 2019), potentially widening educational disparities (Kizilcec et al., 2017). Moreover, most research on MOOCs has typically focused on analyzing data with limited generalization possibilities, such as single MOOCs (Breslow et al., 2013), several courses from a single provider (Chuang & Ho, 2016), or literature reviews that include data from different providers but where the analyses were performed using different methodologies (Veletsianos & Shepherdson, 2016). Therefore, there has been rather limited research on large-scale cross-platform MOOC studies that can provide an overall picture of the actual MOOC ecosystem trends.

Besides, the spotlight of the research and public media has been centered on the main global players, ignoring the growth of the regional initiatives, without offering a broader understanding of the diverse ecosystem of MOOC providers. For example, one of the latest posts from the New York Times indicating that MOOCs were booming during 2020 (Lohr, 2020), mentions only edX, Coursera, and Udacity as MOOC providers. However, a complete list of active MOOC providers compiled by Class Central reports over 35 large MOOC providers distributed around the world (Shah & Pickard, 2019). This list includes global English-speaking MOOC providers like the ones mentioned before, but also regional initiatives like XuetangX in China, Edraak in Jordan, or openHPI in Germany, to name some examples across the many MOOC providers focused on specific regional populations. These regional initiatives have received significantly less attention and remain almost unexplored from a research perspective.

Multiple studies have reported the important influence that cultural factors and language can have when designing online learning (Aydin & Kayabaş, 2018; Bayeck & Choi, 2018; Grandon et al., 2005; Liu et al., 2016). However, there has been scarce research that looks into the impact that the language of instruction, cultural background, or a localized course design can have on learners' educational experiences across different MOOC providers. Therefore, the MOOC movement has not been able to investigate the world's cultural and linguistic diversity, and the research community has not sufficiently attended to the effects of this heterogeneity on global education (Hunt & Tickner, 2015). This is an important gap that needs to be addressed. This study takes a significant step in this direction, advancing current research on global education by providing a better understanding of how online learning patterns are affected by their regional context (Drachsler & Kalz, 2016).

In order to analyze this issue, we will apply large-scale learning analytics. Previous work by Shum introduced three levels where learning analytics can have an impact, the macro, the meso, and the micro (Buckingham Shum, 2012). When these three levels are applied to MOOCs, the micro-level focuses on a single course, the meso-level on a set of MOOCs, and the macro-level provides analytics that informs the whole community (Drachsler & Kalz, 2016). Therefore, we situate our work at the macro level, providing high level insights across MOOC providers, and exemplifying the importance of developing collective enterprises to share large scale data and reboot MOOC research (Reich, 2015). To accomplish this, we formed a research partnership with 15 MOOC providers around the world to create a joint dataset that includes over six thousand MOOCs, more than eight million learners, and over 15 million enrollments. We also conducted a common survey across six of these MOOC providers to understand the preferences to enroll in a platform and their perceptions about their learning experience (N = 10,899). Our sample of providers includes those that have a clear global focus, such as MITx or HarvardX on edX, or FutureLearn, and others that focus on specific regions, such as Edraak in the Arab world or XuetangX in China. From the large-scale analysis that we have conducted, we will address the following research questions that will help signal the importance of locality in online learning:

RQ1. What are the differences between the demographics of specific geographical populations of learners across global and regional MOOC providers?

RQ2. What are the most important priorities of learners when enrolling in global and regional MOOC providers?

RQ3. What are the perceptions that learners experience studying in global and regional MOOC providers?

The remainder of the paper is organized as follows: Section 2 presents related work in the area of learning analytics and studies focused on the effect of learning and culture in learning. Section 3 describes the methodology and materials used for this research. Section 4 presents the results for each one of the research questions, and Section 5 discusses them addressing the implications, limitations, and future work directions. We finalize in Section 6 with conclusions.

2. Related work

2.1. Learning analytics

Learning analytics is an interdisciplinary field that combines three main disciplines: (1) learning sciences, which provide models and theories on how we learn and teach; (2) data sciences such as mathematics and statistics, which provide the methods and tools to analyze data scientifically; and (3) computer sciences, which provide the tools and methods for designing innovative and human-centered ways for supporting learning (Society of Learning Analytics Research (SoLAR) (2021); Archer & Prinsloo, 2020; Romero & Ventura, 2020). Learning analytics has been defined in the literature differently depending on the context or research area, but most researchers agree with the definition coined at the International Conference on Learning Analytics and Knowledge (LAK 2011 in Banff, Canada) as the "measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens, 2013). That is, and according to this definition, learning analytics is a research field that puts at the service of education the most advanced data analytics techniques to better understand how learning occurs and how to support it. Today, learning analytics is an established field of research and practice that gains importance every year (Brown et al., 2020). This is due, in part, to the increasing availability of educational data in higher education institutions collected through Learning Management Systems (LMS) and MOOC platforms; and, in part, to the growing interest of institutions in improving their decision-making processes and make better sense of complex situations involving social and cultural aspects (Buckingham Shum, 2012).

Learning analytics has been especially important for the study of MOOCs and to expand the research in this area (İnan & Ebner, 2020). MOOCs are "gold sources" that generate massive amounts of learners' data, which are diverse and heterogeneous, collected in different settings, and therefore hold a strong potential to apply analytics techniques and methods for understanding how learning occurs. As referenced above, current research in learning analytics and MOOCs can be organized into the three levels of impact proposed by Buckingham Shum (2012): (1) micro-level, which analyzes fine-grained data from individual learners to support them with personalized feedback or provide interventions adjusted to their needs; (2) meso-level, which operates at the institutional level and uses integrated data to optimize institutional processes by supporting transparent decision-making and helping leaders on determining their strategies; and (3) macro-level, which combines data from different institutions and sources to inform the community and transform systems, models or pedagogical approaches.

When applied to MOOCs Drachsler and Kalz (2016), micro-level learning analytics studies are those which focus on analyzing learner behavior on a single course to, for example, to detect students at risk Dalipi et al. (2018); Na and Tasir (2017), or to provide them with personalized dashboards for supporting their study sessions Davis et al. (2016); Jivet et al. (2018); Pérez-Álvarez et al. (2020). Meso-level learning analytics are those that combine data from different courses from the same provider to understand, for example, the characteristics of the learners' population Cagiltay et al. (2020) or the impact of a particular initiative based on MOOCs Hernández et al. (2018). Finally, macro-level learning analytics focuses on the analysis of MOOCs from different providers, countries, and characteristics to inform broader interests of the different scientific and practitioner communities. Examples of macro-level learning analytics studies applied to MOOCs propose interventions for closing achievement gaps Kizilcec et al. (2017) or studies combining data from international and local MOOC providers to understand the impact of these courses in local learners from different cultures Ruipérez-Valiente, Jenner, et al. (2020).

In this study, we benefit from this prior work and recent advances in learning analytics to run a macro-level large-scale study using data from different MOOC providers, that contain courses in different languages and areas, and data from learners from all over the world.

2.2. Language and culture in learning

During the last decades, and particularly since online learning has started to implement more globalized approaches, several researchers have examined the influence of students' native language and cultural background on their motivation and success in learning. Marsh et al. (2002) delineated that instruction in a second language had substantial negative effects on academic achievement. Sulkowski and Deakin (2009) stated that international students in UK higher education institutions are challenged by language comprehension, unfamiliar approaches to teaching and learning, and conforming to the norms of interpersonal conduct. Research in the context of MOOCs has shown that non-native English students interact with videos differently than native English speakers (Uchidiuno, Koedinger, et al., 2018). Therefore, most research seems to convey that a one-size-fits-all approach is not successful for non-native English speakers, and learning experiences need to be adapted based on the needs and motivations of these participants (Uchidiuno, Ogan, et al., 2018).

Students whose native language differs from the course language have to master not only the actual course contents, but also a second language. Therefore, the scarce availability of non-English courses constitutes an additional obstacle towards reaching the goal of closing digital gaps. Several attempts have been made to introduce multilingual MOOCs, but despite the availability of automated transcription and translation software, this has not been a sustainable endeavor thus far. The European Commission has funded several projects in this direction. For example, the *EMMA* project [1] built a platform to provide multi-lingual European MOOCs. During the time

[1] https://project.europeanmoocs.eu/.

it was publicly funded, several multi-lingual MOOCs were produced and published. Currently, the platform is employed as a vehicle to publish (English-only) MOOCs in the context of another project. Another example of a European project is *traMOOC*, also funded by the European Commission, which is set out to provide automated translations for video transcripts and other MOOC components. However, even though state-of-the-art machine learning approaches have been employed (Castilho et al., 2017), more than a third of the participants found the transcripts hard to understand (Sosoni & Stasimioti, 2016).

The overall task of providing such subtitles for videos is still very work intensive as there are many time-consuming steps in-between, such as quality assurance or testing. Although videos are the major educational component of many MOOCs, merely providing subtitles is by far not enough. Quizzes and exams, for example, require even a more careful translation to ensure that there are no ambiguities. Another very important aspect that needs to be covered is the translation of discussion forum content, which to the best of our knowledge, currently, is not supported by any MOOC platform. Particularly in the context of forum discussions, cultural differences enter the scene as an additional factor next to mere translations. Saville-Troike (2003) pointed out that shared knowledge and skills are a key concept for a contextually appropriate interpretation of language in a community. Ogan et al. (2015) detected cultural differences in off-task behaviors, help-seeking, and collaboration in MOOCs. Liu et al. (2016) found that MOOC learners are more likely to interact with forum peers within their own cultural group out of a desire to eliminate language barriers through communication with someone of the same language group. Huang et al. (2019) examined different aspects of the influence of cultural values on technology adoption. Focusing on two countries in the Latin-European and the Confucian-Asian cluster (Spain and China), they compared university teachers' information and technology (ICT) acceptance (Huang et al., 2021) and students' use of mobile technologies for learning (Huang et al., 2020).

Given that language and cultural differences can constitute obstacles in a student's academic achievement in global MOOCs, regional MOOC providers that offer courses in local languages to learners within a similar cultural group, seem to be a promising alternative.

The question remains, however, if the global, multi-cultural nature of these MOOCs should not only be seen as an obstacle but also as a chance to overcome regional silos. In its extreme this "cultural segregation" resembles the experiment of gender segregated education to a certain extent, which was a quite popular approach several years ago but lead to more gender-stereotypical beliefs, biases, attitudes, and behaviors (Fabes et al., 2013, 2015). As part of this study, we also conducted a large-scale survey across MOOC providers to explore the preferences and experiences of learners in global and regional MOOC providers.

3. Materials and methods

The section is divided into several parts. First, we explain our macro MOOC learning analytics framing and second, the methodology that we have pursued. Third, we describe the context of the MOOC providers. Fourth, we detail the design of the survey that we have implemented, and finally, we provide an overview of the final data collection. Following open science best practices to improve educational research (van der Zee & Reich, 2018), we have generated an Open Science Framework (OSF) project with detailed extra information, figures, documents, data, and code that can greatly improve the reproducibility and reusability of this research (see (Ruipérez-Valiente, 2021)).

3.1. Macro MOOC learning analytics framing

We have framed our study utilizing the MOLAC framework (Drachsler & Kalz, 2016), where the overarching goals of this study fall at the macro learning analytics level of the framework. We now depict the different phases conducted as specified by the MOLAC framework.

First, we have worked at the meso level to gather a curriculum of MOOCs from 15 different MOOC providers. Next, we have applied

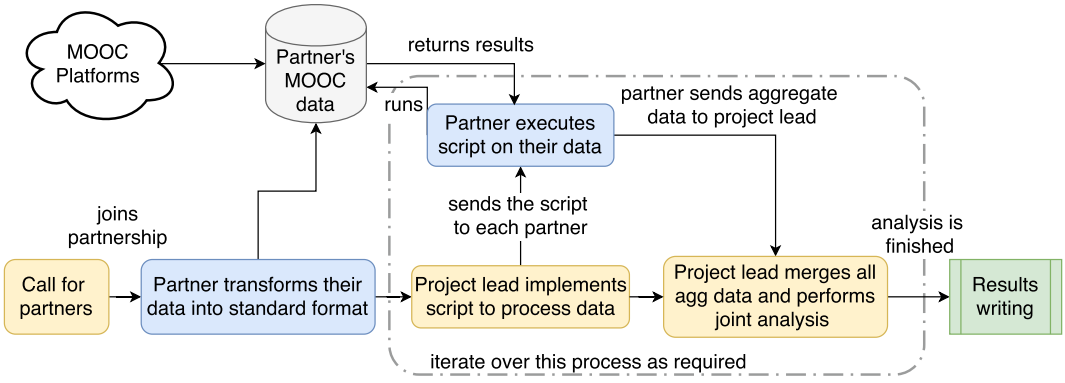


Fig. 1. This figure presents an overview of the methodology that we have conducted to replicate the analysis across different MOOC providers. Blue indicates work executed by each partner separately, work conducted by the project lead, and green collaborative work. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

meta-data standards so that we enable interoperability between these different MOOC platforms. Then, we have applied the learning analytics methods in each one of the datasets separately to avoid sharing data from individual learners. Finally, we have performed data sharing of aggregate data that does not imply any privacy threat.

Then, this meso level output is the one that provides the joint dataset to build the macro level MOOC directory. It is important to clarify that this joint dataset incorporates the meta-data standard which enables cross-platform interoperability on the data repository. The following phase is to perform the macro cross-institutional learning analytics using the full MOOC directory, aiming to respond the three RQs. Finally, the final products of this study will be a set of guidelines for design and technology innovations based on the macro level analyses. Moreover, we also plan to make the resulting aggregate dataset and scripts available for other researchers.

Next, we explain how we have operationalized and implemented these ideas in the following subsections.

3.2. Methodology

There are two main challenges in conducting large scale analytics replicated across different educational institutions and platforms. The first challenge is regarding the privacy concerns of exchanging student data that contain personally-identifiable information at a student-level, especially in those cases where anonymization might not be enough (Daries et al., 2014). The second challenge is regarding the replication of exactly the same analysis in different MOOC providers that might have different data models.

We envisioned a replication methodology that would greatly alleviate both the logistical and privacy concerns and that would also allow us to perform an "apples-to-apples" comparison in datasets coming from different environments (Ruiperez-Valiente et al., 2019). Fig. 1 presents an overview of this methodology.

We operationalized the process to accomplish this goal in the following six steps:

1. **Call for partners**: The initial phase entangled contacting known colleagues and contacts of colleagues to gather a group of researchers with access to a large portfolio of MOOC data (normally from an entire institution), and they should have an interest in performing large scale analytics in the context of global and regional MOOCs. The original partnership proposal is available in the OSF project (Ruipérez-Valiente, 2021).
2. **Establishing a common data format**: We established a common data format for all the datasets in order to avoid exchanging data and to replicate exactly the same analysis in each MOOC provider. They were asked to provide a person-course dataset, where we have a row with the information of the student per each course enrollment, and a course dataset with metadata about each one of the courses that they offered. This dataset contains personal information at a student level, and therefore it is never shared with others. The specific fields with their descriptions are available in the OSF project (Ruipérez-Valiente, 2021).
3. **Feature engineering for data aggregation**: In this step, the project lead implements a Jupyter notebook that will perform the feature engineering to aggregate the dataset of each MOOC provider, and outputs an aggregated dataset in CSV files that do not contain any identifiable information at a student level, and thus are not sensitive. The project lead sends this Jupyter notebook to each partner, they execute the script on their MOOC datasets separately, and finally, each partner sends this aggregated dataset to the project lead. The Jupyter notebook that has been executed by each partner is available in the OSF project (Ruipérez-Valiente, 2021).
4. **Conducting the joint data analysis**: The project lead receives the aggregated datasets from each partner and performs the joint data analysis that is the basis of the study. This phase needs to merge the different aggregated datasets from each provider to perform a joint analysis. It also includes a data cleaning phase to make sure that the labels and values are comparable across MOOC providers. This is performed using a set of scripts written in R, which are available in the OSF project (Ruipérez-Valiente, 2021).
5. **Collaborative interpretation of results**: The final output of the last phase is a set of joint datasets and visualizations with data from all the MOOC providers. These materials are available in the OSF project (Ruipérez-Valiente, 2021). We collaboratively interpret the results so that each partner can bring to the table their knowledge of institutions to help contextualize the data.
6. **Launching a common survey**: The final step was to launch a common survey across some of the MOOC providers. This survey was focused on their enrollment preferences and learning perceptions when taking MOOCs on global and regional platforms.

3.3. Description and context of the MOOC providers

We provide next a brief description of the context from each of the MOOC providers that are involved in this study.

- **MITx** and **HarvardX** (abbreviated as MITxHx): The two original partners in the edX consortium. The majority of courses are taught to a global audience in English.
- **FutureLearn**: Founded by the UK Open University, with over 170 partner institutions globally to provide MOOCs, micro-credentials, and degrees. Most courses are in English.
- **CU Boulder**: The University of Colorado Boulder is a large public university in the US teaching MOOCs in Coursera. Most courses are in English.
- **openSAP**: In 2013 the German-based software company SAP launched their platform for enterprise MOOCs. Based on the HPI MOOC Platform with courses mostly in English.
- **OpenWHO**: Courses offered in a variety of languages with an international focus. Based on the HPI MOOC Platform.
- **openHPI**: One of the MOOC pioneers in Europe, since 2012, the platform has offered courses about digital technologies, transformation, and engineering in German and English. Based on the HPI MOOC Platform.

- **mooc.house**: A white-label platform based on the HPI MOOC Platform, where companies and institutions can offer MOOCs under their own branding. Courses are offered in German and English.
- **HEC Paris**: HEC Paris launched its first MOOC in 2013 and now has offered a wide collection of business and management-related online courses. The courses are offered in either French or English. HEC Paris offers its online courses hosted on the Coursera platform.
- **PUC**: Pontificia Universidad Católica de Chile is one of the most prestigious higher education institutions in Latin America. They offer courses on the Coursera platform on various topics and mostly in Spanish.
- **UAMx**: Supported by the Universidad Autónoma de Madrid and hosted by edX, provides courses across various topics and mostly in Spanish.
- **UPValenciaX**: Supported by Universitat Politecnica de Valencia in Spain and hosted by edX, provides a variety of courses in STEM, nearly all of them in Spanish.
- **UPVx**: Another site supported by Universitat Politecnica de Valencia which is hosted in its own Open edX instance. Focuses on local topics for the Valencian region and basic STEM courses. Courses are in Catalan (Valencian) and Spanish.
- **Edraak**: It was founded in 2013 by the Queen Rania Foundation for Education and Development to serve Arabic speakers and learners in the Arab world. Edraak hosts courses in Arabic on its locally adapted Open edX platform.
- **XuetangX**: Tsinghua University launched XuetangX, the first MOOC platform in China in 2013. With the goal of integrating and sharing global high-quality MOOCs since the beginning, XuetangX has already created over three thousand online courses from first-class universities both in China and all over the world. The platform has attracted over 63 million learners, with a total course registration number of up to 228 million. In our data sample, we only have courses from Tsinghua University hosted in XuetangX.

These providers target different populations either more globally, regionally, or locally. Some of them host their courses in a platform with many other providers (such as edX or Coursera), or they host them in their own instance (such as openHPI or UPVx). There is a broad spectrum of possibilities, but we find that these three are the predominant ones:

- **Global**: Providers that offer a variety of courses predominately in English to a global and international audience, within our data collection we have MITx, HarvardX, FutureLearn, CU Boulder, openSAP, and OpenWHO.
- **Mixed focus**: Providers that offer courses in a non-English local language, but also in English. Therefore, they have a regional focus, but they also offer the possibility to international students to take some of their courses in English. They might host the courses on their own platform or partner with a global platform like edX or Coursera to host them. In our study we have UPValenciaX and UAMx offering courses mainly in Spanish on edX, PUC offering courses in Spanish on Coursera, HEC Paris offering courses mainly in French on Coursera, and openHPI and mooc.house offering courses mainly in German hosted on their own platform.
- **Regional**: Providers that manage their own MOOC platform and offer courses predominantly in a non-English language focusing on a regional or local population, in our scenario we find UPVx, Edraak, and XuetangX.

3.4. Survey design

The survey is centered on the two aspects that are the basis of RQs 2 and 3. The first one is related to the preferences and priorities to enroll in a MOOC platform (e.g., the local language, courses that match interest, among others). The second one is related to the perceptions that they experience when learning in MOOC platforms (e.g., similar cultural background of instructors, adequate course difficulty, similar pedagogy, among others). The survey items have been based on successful surveys previously ran on MOOC contexts (Chuang & Ho, 2016; Ruipérez-Valiente, Halawa, et al., 2020), and only minor adjustments and few new items were needed. Therefore, we can assume the survey instrument as previously validated.

The survey was launched in six MOOC providers and we specifically ask students if they have previously registered to any global English-speaking platform, and if that is the case, we ask the same questions for both global and regional MOOC platforms. Therefore, this allows us to compare both the preferences to enroll in a global and regional platform, and the perceptions when learning in a global or regional platform, which is the main focus of the survey. Finally, we additionally requested several demographics and other data.

Each partner was responsible for translating the survey to the local language of each one of the MOOC providers (German, Spanish, Arabic, and Chinese) and managing all the implementation logistics to send the survey to a randomized sample of their learners. We launched the survey in openSAP, openHPI, UPVx, UPValenciaX, Edraak, and XuetangX. We launched the survey in these providers based on our focus on regional providers (the last five of them can be considered to have a regional focus, while openSAP has a more global focus on professionals) and availability. The sampling strategy of the learners that we followed was the same in all the MOOC providers. Each one of the partners selected a sample of the learners that had previously registered for a course in their platform and were accepting emails. The survey was then submitted via email inviting the learner to participate in this international research that is exploring preferences and perceptions of MOOC learners. Therefore, all students had previous experience participating in a MOOC on that platform. The final data sample size and the response rate varied significantly from platform to platform, but since learners were sampled in a random way, we do not expect biases. The full survey is available as part of the OSF project (see (Ruipérez-Valiente, 2021)).

3.5. Data collection

Based on the previously depicted methodology, we collected datasets from each one of the providers regarding the demographics of

learners (country of origin, age, level of education, year of birth, and gender), and regarding their activity with the course (if the student viewed the course and completed it). We additionally collected metadata from each one of the courses, such as the language, and the topic. Fig. 2 shows the distribution of the language with MOOCs in English (*en*), Spanish (*es*), Chinese (*zh*), Arabic (*ar*), French (*fr*), German (*de*), Catalan (*ca*), Bulgarian (*bg*), Japanese (*ja*), and Italian (*it*) for each one of the providers.

Additionally, we categorized courses into four broad areas of knowledge defined by previous work (Chuang & Ho, 2016): Government, Health, and Social Science (GHSS), Humanities, History, Design, Religion, and Education (HHRDE), Computer Science (CS), and Science, Technology, Engineering and Mathematics (STEM). The distribution of the number of courses in each category for each provider is available in Fig. 3.

The topics, format of the platforms, length, and instructional design can vary greatly from one course to another, generating a high heterogeneity between this large sample of courses. Previous work developed by García-Peñalvo et al. (2018) classified MOOCs in different models, such as those focused on connectivist approaches (cMOOCs), content-centric approaches (xMOOCs), hybrid approaches (hMOOCs), or adaptive hybrid approaches (ahMOOC). In our case scenario, the large majority of the MOOCs are content-centric xMOOCs, meaning that instructors develop video-lectures, slides, exercises, and other contents for students to consume. Many of these courses might have forums or other social/collaborative tools, but these are not the focus of knowledge creation. Some of these courses are taken on a fixed-schedule (each week a unit is released and there are multiple due dates along the course) and others are offered on a self-paced format (all contents are released since the beginning and learners have a final course due date to complete everything). In terms of topics, Fig. 3 showed the high diversity across platforms, where some of them have a focus on STEM or CS topics like openSAP or openHPI, others have a focus on health like OpenWHO, or on business like HEC Paris, and then many other platforms have a good proportion of courses in all course categories. Finally, the length of the courses is also diverse, with some of them taking place over just two or three weeks, and others having a length of up to 12 weeks.

Full details of the process to collect the data collection are available in the OSF project (Ruipérez-Valiente, 2021), and the specific numbers of learners, courses, and survey respondents per provider are available in Table 1.

3.6. Analyses

Once the aggregated data have been put together in an interoperable format, we can perform the macro cross-institutional learning analytics to respond to the RQs, mostly by providing exploratory analyses using descriptive statistics and visualizations.

For RQ1, we focus on comparing specific geographical sub-populations across the different MOOC providers within our data collection. The following selected sub-populations are based on the country where each one of the providers operates, as well as the predominant language of their courses. We have selected Hispanic (all Spanish-speaking countries [2]), Arabic (all Arab world countries [3]), Chinese, German, and French sub-populations. We will present stacked bar charts of the level of education, gender, and age for each one of the aforementioned sub-populations in each one of the MOOC providers separated also by the course language. This will allow us to discern how the demographics of these sub-populations change from one provider to another.

For RQ2-3, we focus on comparing the aggregate survey responses in terms of preferences and perceptions in global and regional settings across the six MOOC platforms. To do so, we will use bar charts for each provider separately, that will compare the preferences and perceptions of their learners when studying in global vs regional providers. This will allow us to see those dimensions where learners really find differences between the two MOOC scopes.

4. Results

4.1. RQ1. Distribution of demographics across global and regional MOOC providers

Fig. 4 shows a stacked bar chart with the ten most representative countries per provider. The results show that global providers recruit about 30% of their audience from their home country: MITx and HarvardX attract 30% learners from the USA, while FutureLearn has 30% of learners from the UK, and CU Boulder has 23% from the USA. These proportions are much higher among providers that have a more regional emphasis and local languages, for example, 70% of students from openHPI and mooc.house are German, 76% of learners from UPVx are Spanish, and Edraak and XuetangX have almost 100% Arab and Chinese learners, respectively. This pattern is also evident in absolute terms, for example, based on our data and Class Central reports we find that up until 2018 Edraak had around 1.5 million Arabic learners, and XuetangX 14 million Chinese learners, compared to just 163 thousand Arabic learners, and 100 thousand Chinese learners in all MITx and HarvardX courses. We see more diversity in the demographic distribution in those providers located in global platforms but with a more regional emphasis, such as HEC Paris, PUC, UPValenciaX, and UAMx. We suspect that structural language differences drive some of these adoption patterns; especially in those languages that are structurally different from English.

We look now into the differences in the level of education, gender, and age across the different sub-populations comparing global and regional MOOC providers, and splitting by the course language.

First, in Fig. 5 we see the distribution of level of education. In general, we observe lower levels of education in regional MOOC providers, especially if the courses are taught in the local language. This pattern is very clear with 85% of Edraak's learners and 78% of

[2] https://en.wikipedia.org/wiki/List_of_countries_where_Spanish_is_an_official_language.
[3] https://en.wikipedia.org/wiki/Arab_world.

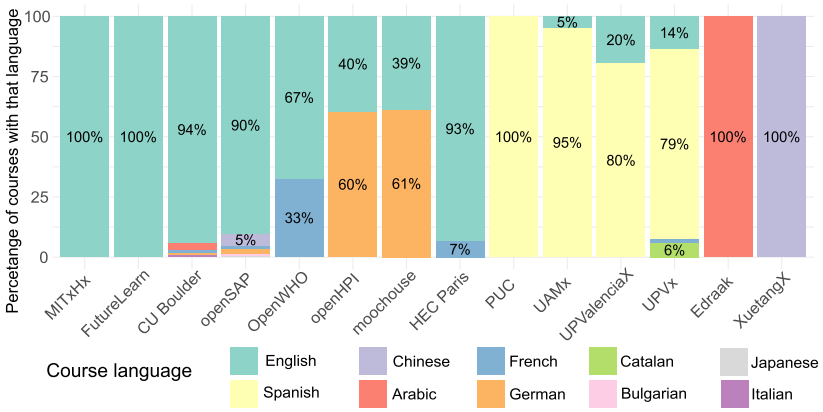


Fig. 2. Boxplots with the distribution of course language for each one of the MOOC providers.

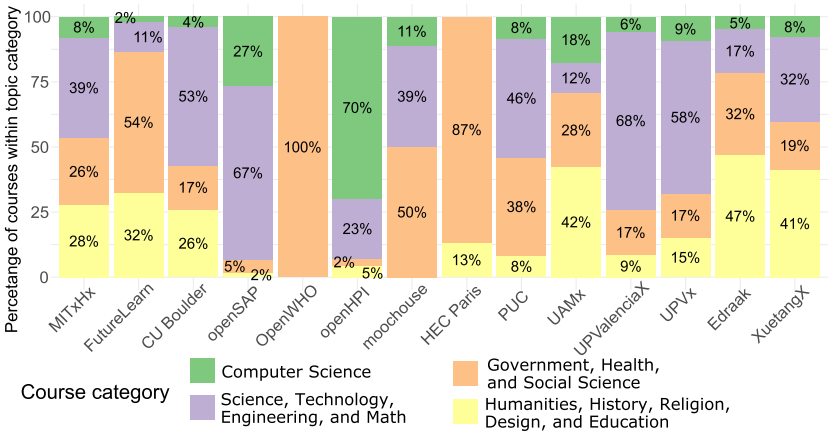


Fig. 3. Boxplot with the distribution of the course category for each one of the MOOC providers.

Table 1
Overview of the data collection with the number of courses, learners, and survey respondents collected from each MOOC provider.

Provider	# MOOC instances	# Learners	# MOOC enrollments	# Survey respondents
MITxHx	552	3.72 million	6.81 million	N/A
FutureLearn	1548	1.16 million	2.06 million	N/A
CU Boulder	111	541 thousand	679 thousand	N/A
openSAP	166	515 thousand	1.64 million	519
OpenWHO	52	36 thousand	59 thousand	N/A
openHPI	43	113 thousand	263 thousand	1038
mooc.house	18	24 thousand	33 thousand	N/A
HEC Paris	33	22 thousand	29 thousand	N/A
PUC	24	140 thousand	179 thousand	N/A
UAMx	90	182 thousand	225 thousand	N/A
UPValenciaX	309	914 thousand	1.2 million	8380
UPVx	182	43 thousand	52 thousand	198
Edraak	228	610 thousand	1.47 million	431
XuetangX	2884	655 thousand	1.03 million	333
Total	6111	8.67 million	15.73 million	10,899

XuetangX's learners having either a bachelor or lower education, compared to the distributions that we see in the global providers, with for example only 41% of the Chinese learners in CU Boulder having a bachelor or lower. Additionally, we also observe differences in those providers that have courses both in English and a local language. For example, we see less educated learners in the Spanish courses of UAM, UPValencia, and UPVx than in their English courses. Therefore, the course language is playing a key role in the level of education of the learners each provider attracts.

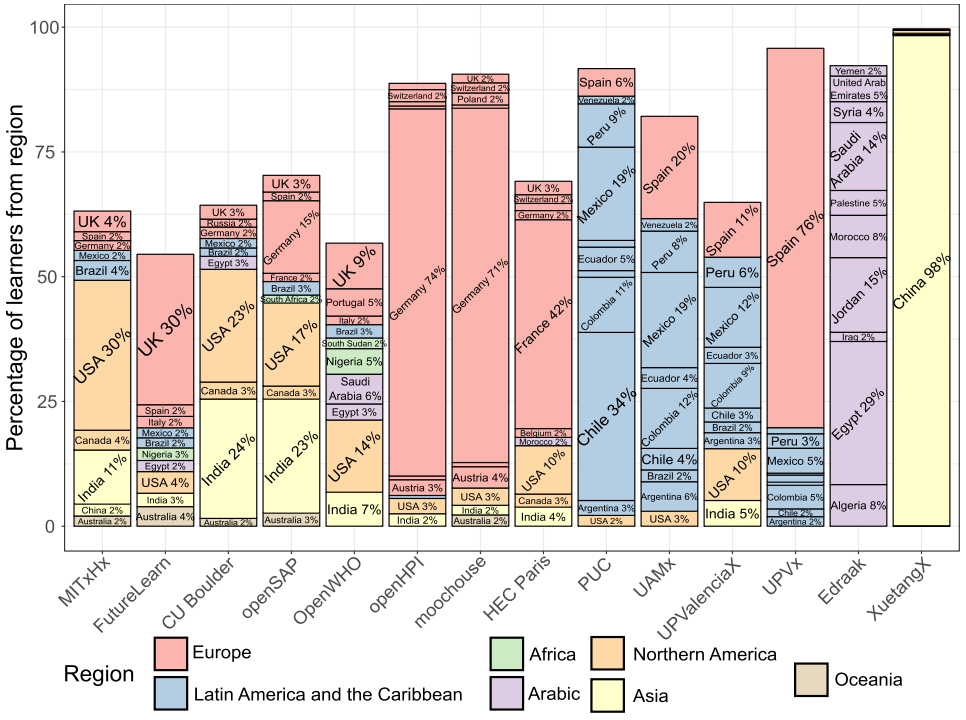


Fig. 4. Stacked bar plot with the top-ten most representative countries in percentage per provider. The color identifies the region of the country. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

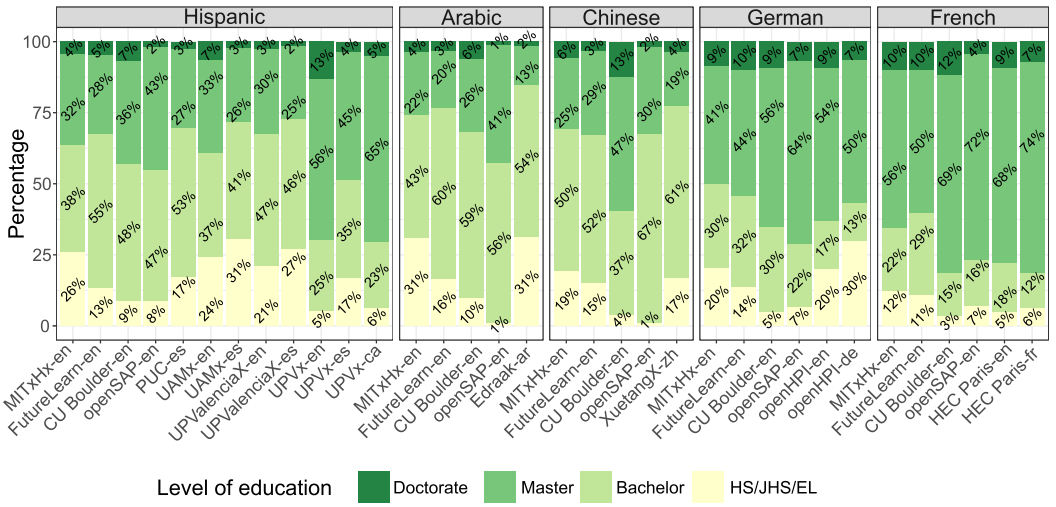


Fig. 5. Distribution of level of education per MOOC provider and course language for each one of the analyzed sub-populations.

Second, in Fig. 6 we see the gender distribution. There is not seem to be a clear pattern difference between global and regional providers, nor a clear influence of the course language within the same MOOC provider. We do observe a significant pattern where the topic of the course has a strong influence on the gender of the attracted participants. In general, we observe that women tend to enroll more in GHSS/HHRDE courses, and men in CS/STEM, this is consistent across the different MOOC providers (see additional plots in the OSF project additional materials (Ruipérez-Valiente, 2021). Therefore, this can intrinsically affect other patterns that we observe, since as we reported each MOOC provider has its own distribution of course types.

Finally, in Fig. 7 we see the age distribution on different buckets per each provider. We do observe some interesting patterns, for example, the Arabic and Chinese populations of learners are younger than the European ones. Moreover, in all sub-populations (except

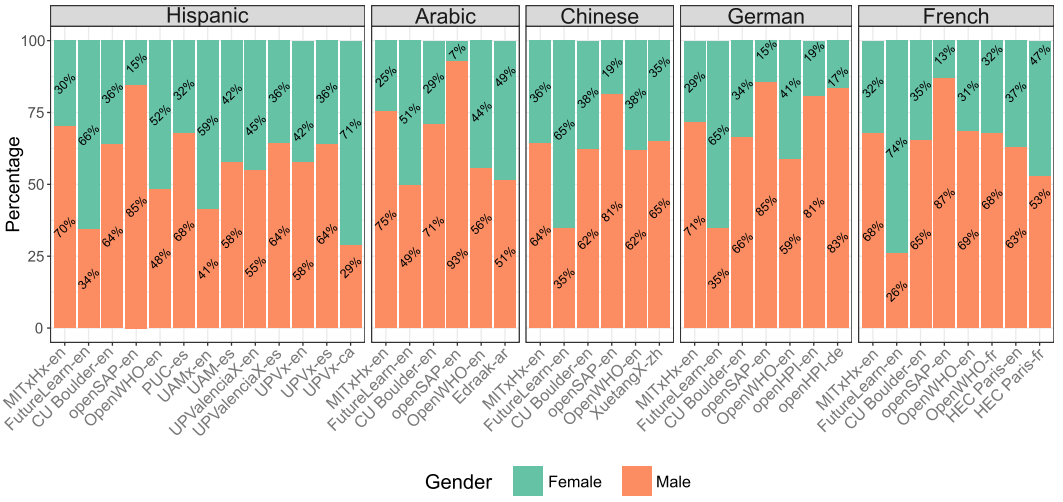


Fig. 6. Distribution of gender per MOOC provider and course language for each one of the analyzed sub-populations.

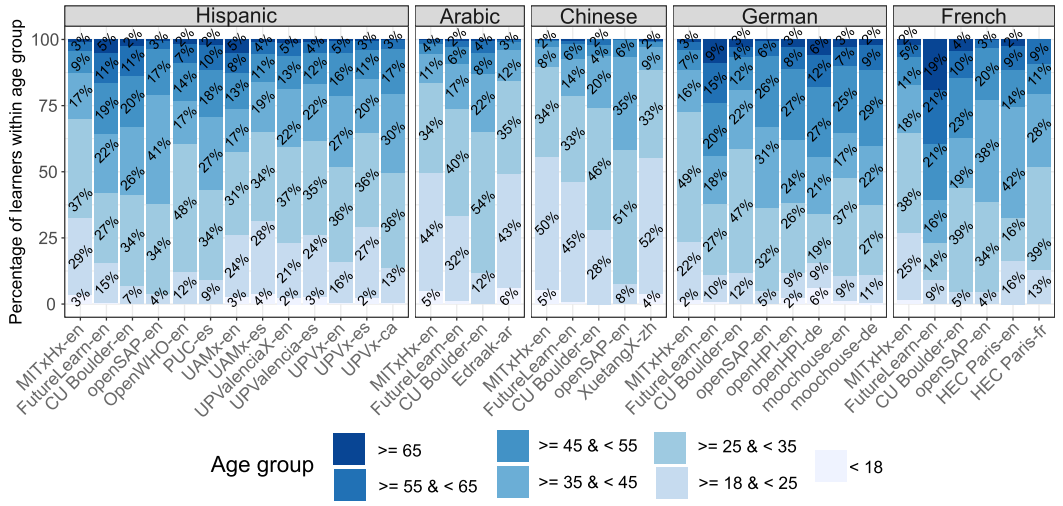


Fig. 7. Distribution of age per MOOC provider and course language for each one of the analyzed sub-populations.

German), courses in local languages attract younger learners. However, we do not observe consistent differences in terms of age across the different sub-populations in global and regional providers, and the differences seem to emerge more locally at a provider level.

4.2. RQ2. Importance of reasons for enrollment in global and regional MOOC providers

We find the results from the survey items associated with this research question in Fig. 8. The barplot compares the preferences to enroll in a global platform versus a regional platform for each survey items separately, and the results of each provider are presented in a different sub-plot. See the full survey statements in the OSF project repository (Ruipérez-Valiente, 2021).

The results show that the primary preference when enrolling in a platform for learners across all providers is that courses "Match my interests," and this is consistent for both global and regional settings across the different MOOC providers. There are other items also highly rated consistently, such as "Quality of courses" and "Learning opportunities." Therefore, their primary goal is on the learning side. Additionally, other items that have a higher variability; for example, on the social side we find that "Connecting with others" or "Online Community," are much more highly rated on Edraak and XuetangX than on openSAP and OpenHPI. Additionally, we also see that for example in UPValenciaX and UPVx, they value more this social dimension in local settings than in global ones, indicating that they can connect more easily with people in their same geographical location.

Finally, the item "Language" that asked respondents about enrolling in courses in their local language, also has high variability. We find that on average learners from UPValenciaX, and Edraak find this issue important, but that there are no differences between

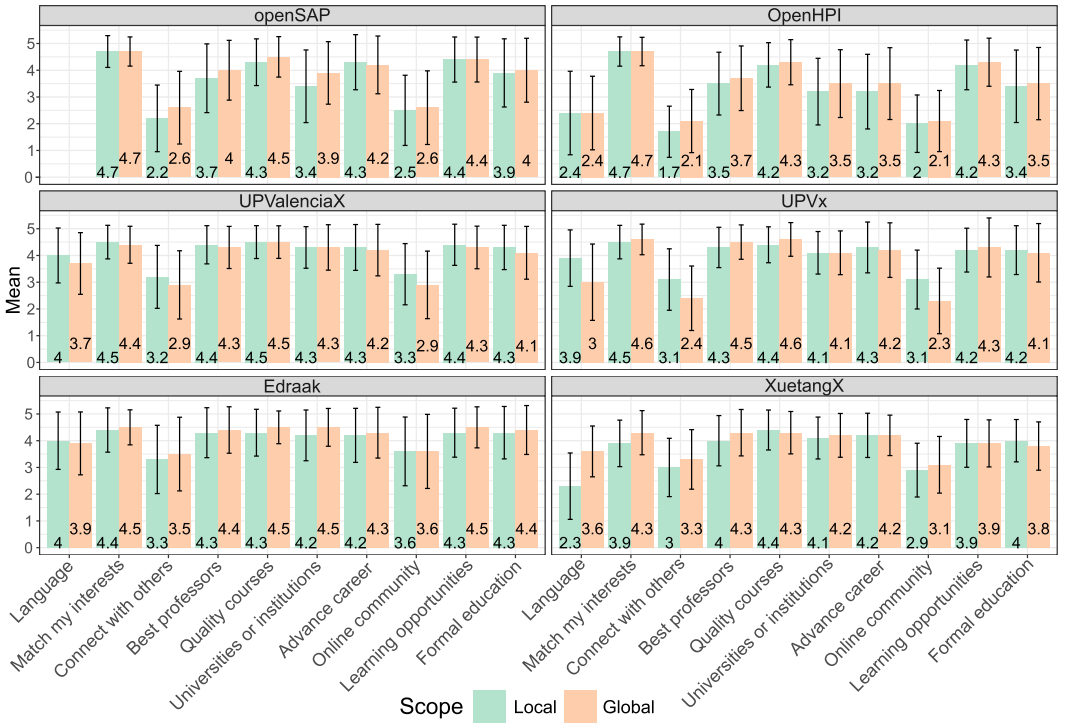


Fig. 8. Average values of the survey results on the preferences to enroll in a platform per provider and question and split by the global or regional scope of the platform.

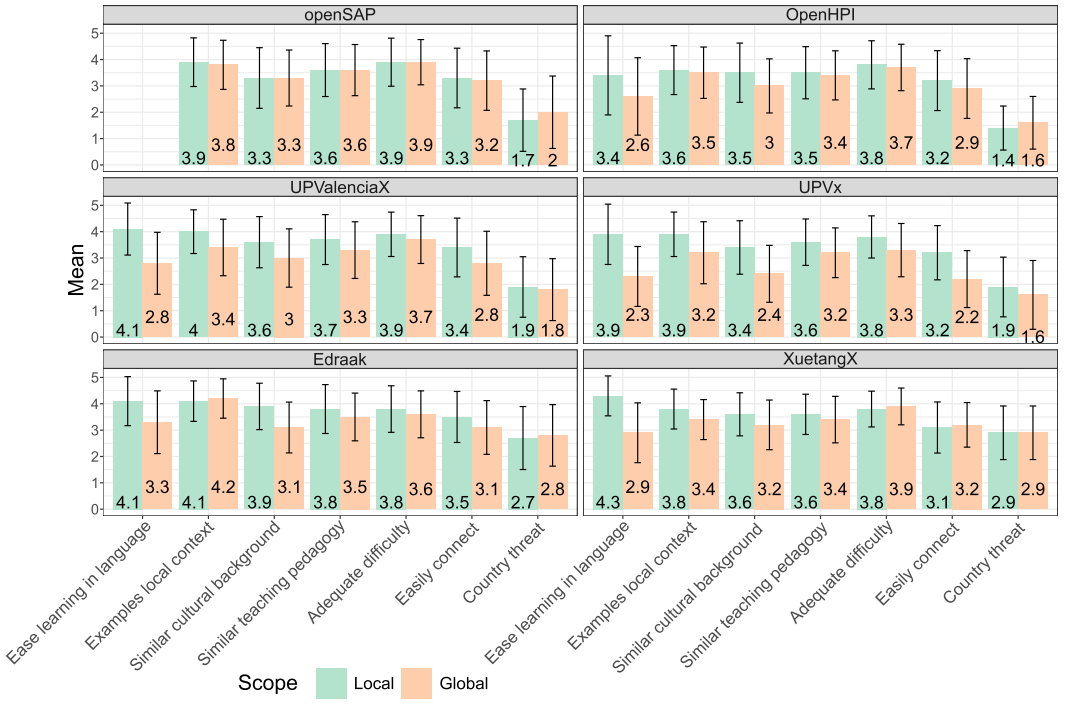


Fig. 9. Average values of the survey results on the experienced perceptions when learning in a platform per provider and question and split by the global or local scope.

regional and global providers, whereas learners from openHPI find this issue non-important, potentially because the German population has a high average English proficiency. Moreover, learners from UPVx preferred taking courses in their local language, but learners from XuetangX prefer studying in global settings, potentially to practice their English skills.

Therefore, all these results suggest that learners that enroll either in a global or regional MOOC provider, do that because those platforms are the ones that satisfy their needs the most. Additionally, it also means that there are other educational and social aspects that learners care about which are dependent on the scope of the provider, and that sub-populations might care about these aspects differently. All of these are important features that should be taken into account to satisfy learners' needs.

4.3. RQ3. Perceptions when learning in global and regional MOOC providers

Analogously to the previous plot, we find the survey results in terms of the perceptions that students experience when learning in regional versus global platforms in Fig. 9. The first key finding which is consistent across the five regional providers is that learners can learn more easily in their local language than in English. There are also other multiple items on cultural aspects, such as "Examples of local context," or "Similar cultural background," that have higher ratings in local context than in global ones. These gaps are different from provider to provider, for example, we see a difference of one point in "Easily connect with others" for UPVx learners, but this gap is non-existent for XuetangX learners. Multiple cultural and social aspects might be at play here.

A positive result is that there is no gap in "Country threat" in global and regional settings, however, we do see that the average value of "Country threat" is in general much higher for Arabic and Chinese populations than for the other ones. Therefore, learners from specific countries might be afraid of the conclusions that others might draw based on their country of origin, and this applies to both global and regional contexts.

Therefore, these results indicate that students feel more comfortable and at ease when learning in a local setting in several aspects, but especially those related to using their local language and having instructors with a similar cultural background.

5. Discussion

This study sought to analyze the differences in terms of demographics, preferences, and perceptions of learners across global and regional MOOC providers. Overall, the resulting trends suggest that regional MOOC providers are doing a better job catering to their local learners than the global, elite providers. This analysis has only been possible by pursuing a macro MOOC learning analytics initiative. As Drachsler and Kalz (2016) indicated, this large scale multi institutional approach allows us to provide insights for the whole community that go beyond contextual factors. With data from over eight million learners and six thousand MOOCs distributed across nine countries, this study represents one of the largest macro learning analytics efforts in the literature.

Our results have shown an important impact of locality in the learners' population that each provider attracts. Additionally, we have seen that for the sub-populations that we have explored in each regional platform, the demographics that we observe are slightly more inclusive, reaching underserved learners in global providers. Learners that responded to the survey reported that the main reason to enroll in a platform, whether a global or a regional one, is that courses match their interests. Therefore, if there are large amounts of local learners in regional MOOC providers, this might imply that the courses are more appealing to them. Moreover, a nuance to take into account is that learners also reported feeling culturally more at ease when studying in regional MOOC providers. The cherry on top is the local language, while all learners reported that they can learn more easily in their local language than in English, some learners did not have a strong preference to enroll in courses instructed in their local language or English, others preferred their local language, and others English (perhaps to practice). This indicates that some specific sub-populations around the world are more in need to have courses produced (or translated) into their local languages than others. Moreover, the topic and target group of the courses also have a strong impact on the population that they attract. We found that is consistent across different MOOC providers that women tend to enroll more in GHSS/HHRDE courses, and men in CS/STEM. Therefore, for platforms like openSAP or openHPI where 90% of their courses are CS/STEM, is normal to find a gender imbalance with a majority of the male population. On the other side, a platform like OpenWHO or FutureLearn that the majority of their courses fall within GHSS/HHRDE, has a higher proportion of female learners than male ones.

There have been several enablers that have significantly contributed to developing this study. The first one was a successful outcome from the initial phase looking for partners to join the study. This might not seem obvious but this phase was critical to gather a meaningful MOOC directory and specific potential MOOC providers were targeted with an outline of the research project. While most of them regarded this study as interesting and valuable, many indicated that they did not have the human resources to collaborate or were not responsive after expressing their interest. Therefore, this was a critical phase where the study could have failed without even starting. Another key enabler was establishing a meta-data standard that would enable interoperability between the entire MOOC directory. This facilitated conducting exactly the same analyses without performing individual learner data sharing. However, the common meta-data standard is rather limited, and this multi-platform learning analytics would be greatly facilitated if platforms implemented their logs using established interoperable standards like xAPI (The Advanced Distributed Learning Initiative, 2013). Furthermore, the similarities between different MOOC platforms also helped to make this study feasible, even though we had to reach common agreements regarding some variable definitions. This process might be more complex in other virtual learning environments like educational games or intelligent tutoring systems that are more heterogeneous.

This macro scale approach also has its limitations. For example, we have scarce contextual information about each one of the MOOCs included in the study, and more shallow knowledge about platforms than if we were focused on a small portfolio of MOOCs. Obviously, the current macro learning analytics methodology has not allowed us to analyze the data at a micro level focusing on

individual learners, thus limiting our conclusions in those aspects. Moreover, since the majority of current learning analytics and educational data mining research is more focused at the micro level, we do not have a wide diversity of methods that we can apply at a macro level.

Finally, in addition to the presented macro level analyses, the other products of this study are a set of guidelines for design and technology innovations based on the analyses, that in the future can be implemented at a meso level, therefore increasing the education quality of learners at a micro level. Moreover, we also have shared on a public repository the resulting aggregate multi platform dataset and the scripts used to execute the entire methodology (Ruipérez-Valiente, 2021). More specifically, we discuss the implications of these results in terms of design recommendations (Subsection 5.1), technological implications (Subsection 5.2), and broader impacts, limitations and future work directions (Subsection 5.3). We hope this discussion can serve as a guide for both MOOC providers and researchers interested in advancing the field forward.

5.1. Design recommendations

Based on the results, we discuss some design recommendations that can help produce more inclusive MOOCs.

Consider creating content in additional languages. We hypothesize that the observed adoption patterns are related to the structure of the students' first language; especially in those languages such as Chinese or Arabic with a different origin than Roman and Germanic languages (Tremblay et al., 2016). People want to learn on platforms built to support their own languages and in communities with cultural alignment. There is also a connection between the providers and their primary reliance on regional organizations to develop courses. Many of FutureLearn's course development partners are UK-based organizations, as well as are many edX and Coursera partners which are in the United States. This is similar to more regional providers such as mooc.house with German partners, XuetangX in China, HEC Paris in France, and Edraak in Jordan. Platform use is linked to individual learners' likelihood in platform discoverability, either via educational brands they look for, recognize and trust, and from within what they discover in their own social media and filter bubbles (Pariser, 2011). Therefore, providing courses in other languages could be key for attracting local learners and creating links to regional communities. However, how is the actual impact of designing courses in students' first language in terms of the learning experience? Interdisciplinary research including experts in language could shed some light on open challenges.

Cater cultural diversity. Multiple previous studies have reported how learners from developing regions struggle to succeed in global providers (Kizilcec et al., 2017; Reich & Ruipérez-Valiente, 2019; Ruipérez-Valiente, Halawa, et al., 2020), and thus, regional providers might be also better positioned to facilitate learners achieve their objectives decreasing completion gaps. One potential step forward is to start applying frameworks to dimensionalize the cultural characteristics that are present in these environments, and one prominent model that can be applied is Hofstede (2011). In fact, previous work analyzed MOOC videos produced by three different nationalities (France, USA, and South Korea) applying Hofstede's model, and found significant differences in the cultural dimensions (Bayeck & Choi, 2018). In addition, prior research has shown that localizing the content of a course produced within a particular region to another by simply translating the videos might not be enough for engaging local students. This previous study analyzed the effect of localizing courses in English from MITx to Edraak, by translating all contents and generating subtitles in Arabic (Ruipérez-Valiente, Halawa, et al., 2020). The localized courses had much lower levels of engagement and completion in Edraak, and thus we cannot assume that localizing content has the same effect as producing courses inclusively-designed courses from scratch (Aydin & Kayabaş, 2018). Students need cultural references, codes, and rules that facilitate their learning process.

This prior work suggests that the way in which MOOCs are produced has cultural effects on their contents, which in turn can directly affect learners. These differences help to explain why a small number of global elite MOOC providers are unlikely to monopolize online learning, and there are dimensions where regional providers can compete effectively with English-language providers. Therefore, global MOOC platforms, producers, and instructors need to re-envision how to develop courses that can cater better to the diversity of learners around the world. One alternative that has been explored is to facilitate the existence of several cultural communities within the same MOOCs. The study by Colas et al. (2016) organized a MOOC with seven teams of facilitators that promoted active participation and peer tutoring to groups of learners with different native languages. While the results were promising, it is unknown if this method of instruction can be replicated systematically and on a larger scale. How this type of multi-cultural instructional design can be scaled up? And, what is the effect on learners' attrition rate? These are still open questions that will help MOOCs to adapt to the local practices and really accomplish the global reach that they target (Godwin-Jones, 2014).

Select course topics related to regional challenges. Our results indicate that learners found that courses in regional platforms are more aligned with their interests. We hypothesize that this result might be related to the fact that regional providers offer MOOCs that help competence development for addressing regional challenges and needs. They tend to carry out surveying efforts to better understand their population needs. So, identifying what the regional problems are might help define the course topics that can attract more learners and help local development.

Create content to foster an equitable and gender balanced learning community. Previous work found several course features that are significant predictors of enrollment, one of them was the presence of gender cues (Kizilcec & Kambhampaty, 2020). Therefore, strategic design of certain course features can help attract a more diverse and equitable population of learners. For example, the FOSTWOM project provides toolkits for promoting good practice for gender balance in MOOCs (European Commission's Erasmus+for Higher Education, 2021). Another example is the MOOC STEAM4ALL, which is a MOOC devoted to more inclusive teaching practices in STEAM (Gonzalez et al., 2021). These types of projects could help establish the baseline for attracting a more balanced population, especially in STEM courses. However, more studies along this line are still necessary to understand the impact of these practices. How are these good practices affecting the attrition rate of students of different genders? How are these best practices implemented practically in the course design? These are questions that require further exploration.

There are numerous open questions in terms of design that should be tackled both by conducting more empirical studies to understand the effect that design principles can have on MOOC learners and their outcomes (Oh et al., 2019), but also from a theoretical perspective, proposing theory-driven frameworks for MOOC development that practitioners can utilize to design their MOOCs (Steffens, 2015).

5.2. Technology implications

Technology can play a key role in successfully approaching many of the aforementioned design requirements. Some of these technologies are in a mature stage and could be systematically incorporated into MOOC platforms, others are still in the development phase.

A basic functionality that all MOOC platforms should include is the possibility to have audiovisual contents in several languages and cohort-based learning, providing a simple and transparent functionality to support personalized contents to different cohorts based on learners' preferences. While some previous work in MOOCs has developed adaptive modules that can adapt the exercises (Rosen et al., 2018), this functionality would target a learner-driven personalization of the learning experience based on learners' preferences. For example, it does not make sense for a Spanish learner that enrolls in a MOOC to learn in English and to connect with the wider international community, to be assigned to the Spanish cohort having all the contents translated and to interact with a sub-forum with other Spanish-speaking learners. While the work of Colas et al. (2016) showed a promising example to support several sub-communities in forums, these functionalities were not directly included in the MOOC platform. Moreover, MOOC platforms should also have feedback functionalities that learners can use to provide information about their learning experiences and their preferences in terms of the courses that they would like to take; this would help implement feedback loops for the continuous improvement of the platform and course portfolio. Developing technologies that can empower learners to design their own learning experiences might be key to catering to the current diversity.

One of the main issues to implement these approaches is the escalation in the production costs to generate content in multiple languages or with different instructors, for example. It is therefore key to improving production pipelines that can maximize cost-effectiveness (Santos Espino et al., 2021). Although we have seen in the previous section that simply translating the videos might not be enough in some cases, there are several technologies that can provide support in that sense, such as the use of automatic subtitling with speech recognition (Pérez-Martín et al., 2020), and the translation of the text to other languages. The new text in a different language could also be transformed into new audio, by applying AI-driven text-to-speech approaches (Dao et al., 2021). The advancement of dynamically translated written content and text-to-speech could be one way to help regionalize global MOOC providers/courses, mix cohorts across different native languages, and permit support for various languages all within one platform (i.e. you select the language alongside enrollment). These approaches could also be extended to have peer-assessment and collaborative work implemented in a way that involves only those learners within the same cohort. Furthermore, additional work might be needed from the instructors' side to manage the course, since now there might be multiple contents to support. To alleviate this burden, MOOC platforms should also implement learning analytics solutions that can support the instructor to detect learners' in need and intervene (Xing & Du, 2019), to detect issues in the content via curriculum analytics (Kitto et al., 2020), or recommend threads in the discussion forums that need the intervention of a subject-matter expert (Alrajhi et al., 2020; Wong et al., 2018).

Finally, perhaps the most complex issue to improve via technology is the cultural one. One incoming technology that could play an important role in audiovisual translation to multiple contexts is the research area of deep learning-driven methods to generate synthetic media, as these could be used to create culturally adapted media across multiple regional contexts. The surreal and scary vision of deepfakes (Verdoliva, 2020) could also be used to improve education across contexts. For example, this would allow the possibility for a learner to select the physical appearance, language, and even the accent of the instructor they want teaching in the videos of the MOOC they enrolled to. Another technology that has been advancing at a rapid pace is automated essay scoring (Kumar & Boulanger, 2020), which could perhaps improve the issue of essay assessment of non-native speakers. However, a generalized disadvantage with all these technologies is that on most occasions algorithms are trained using a dataset in English, and they might not perform well in other languages yet. Moreover, empirical research would be needed to assess if these solutions can actually help learners or if they would worsen the learning experience.

5.3. Broader impacts, limitations, and future work directions

A key question that has been within the community is the role that MOOC platforms can have in the higher education system. In that sense, over the last few years we have seen how several MOOC providers operating at a global scale, have been transitioning by raising content paywall barriers to learners and partnering with universities to offer fully online or blended programs (Reich & Ruipérez-Valiente, 2019). In line with the story, it was recently announced the sale of edX to the for-profit company 2U (LeBlanc, Paul, 2021). The business directions of these large global MOOC platforms for-profit are clear, but there is still hope for the regional initiatives to remain truthful to the original promises of MOOCs (Godwin-Jones, 2014). Many of these regional platforms are funded by national governments or non-for-profit foundations that seek to provide a high quality open education to a regional society. However, they might be facing in the future the same sustainability problems that other global MOOC providers faced before them (Rodriguez et al., 2018). Therefore, more systematic research around these regional initiatives is needed, as the current literature remains quite scarce.

Our methodology has exemplified a potential way forward to perform collaborative work with regional MOOC providers in order to generate research that they might not be able to carry out otherwise. Moreover, we have also exemplified the importance of developing

large-scale studies, since some of the trends that are observed in particular courses or even at a platform level, might not be replicated in other contexts. Our framework has been successful at performing a truly large scale collaboration between many partners, while maintaining the privacy of the learners and without sharing student-level data. We encourage other researchers to re-use this framework, and even the code and data (Ruipérez-Valiente, 2021), to further pursue this approach. However, if we really want to scale this approach, we need a technical infrastructure that can centralize all these functionalities and provide an API that researchers can interact with, while at the same time maintaining all the needed features. Perhaps, the most noteworthy initiative in that sense has been the MOOC Replication Framework (MORF) (Gardner et al., 2018), an open-source platform-as-a-service (PaaS) that allows conducting privacy-preserving experiments on MOOC data. However, few partners joined this initiative. This kind of privacy-preserving infrastructure that can allow performing experiments on large scale data has the potential to become a key game-changer, especially within the context of MOOCs since all of these platforms have many commonalities and therefore it is more feasible to implement this kind of aggregator (Ruiperez-Valiente et al., 2019).

Our work also has some limitations that we would like to raise. Much of our results are observational, therefore, we can only establish reasonable hypotheses but no causation. Moreover, given the high heterogeneity and the large number of MOOCs, this study focused on large-scale trends as we currently cannot provide a granular and detailed analysis based on the characteristics of each course and provider. Future work should envisage ways to characterize course and platform features in a way that can scale to large data samples. Therefore, we encourage researchers to further analyze the importance of regional MOOC providers within the MOOC ecosystem and higher education. Moreover, more research exploring inclusive course design approaches and testing their efficacy is needed. Another potential direction is having richer details about the preferences and experiences of learners in global and regional settings by conducting semi-structured interviews or interviews with focus groups. Finally, this work has focused only on learners with previous experience on MOOCs, however, non-adopters of MOOCs might bring complementary views to the ones that we share in this project, and thus we recommend future work to consider this cohort as well.

We consider the line of work conducted in this study of the utmost importance given that online learning is becoming more pervasive, and MOOCs have experienced a large growth with new programs and partnerships with universities over the last years. In a world being transformed by a climate emergency, interrupted schooling from pandemics, extreme weather events, and civil unrest (Reidmiller et al., 2017), remote learning will become more common and increasingly important.

6. Conclusions

To better understand how online learning opportunities are expanding through the MOOC ecosystem, we created a research partnership among 15 different providers from nine countries. We gathered data from over eight million learners, six thousand MOOCs, and we conducted a large-scale survey with more than 10 thousand participants. Our work has highlighted the importance of local education in the global MOOC ecosystem. We find that regional MOOC providers attract a high proportion of local learners, with more inclusive demographic profiles, they match better the interests of their learners, and that the cultural alignment is better than in global MOOC providers. While other studies in MOOCs also reached similar conclusions regarding the importance of cultural and language barriers (especially studies on MOOCs in economically disadvantaged regions) with much smaller sample sizes e.g. (Gupta, 2019; Ma & Lee, 2019; Ruipérez-Valiente, Halawa, et al., 2020), our work has been able to replicate these findings in a large scale sample thanks to a partnership between 15 MOOC providers. Therefore, this work builds further confirmation to compound evidence-based research that highlights the importance of cultural and language factors in global learning. Finally, since our entire methodology, questionnaires, code, and data are publicly available (Ruipérez-Valiente, 2021), future researches can build on top of this work to produce new results based on these resources, or replicate this methodology across other MOOC providers.

In the early days of MOOCs, technology evangelists believed that a small group of elite universities would revolutionize higher education by broadcasting courses from the world's most prestigious universities to all corners of the globe (St. Amant, 2007). But even in online learning, language, culture, and diversity matter enormously to many learners. Regional providers may have much to learn from global institutions about best practices in production and marketing at scale to build sustainable institutions. But global providers have much to learn from their regional colleagues about inclusive design and addressing learner needs. The whole online learning ecosystem would greatly benefit from further conversations and research that address the diversity of MOOC providers.

Credit author statement

José A. Ruipérez-Valiente: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Project administration, Writing – original draft, Funding acquisition. Thomas Staubitz: Data curation, Writing – original draft, Writing – review & editing. Matt Jenner: Data curation, Writing – review & editing. Sherif Halawa: Data curation, Writing – review & editing. Jiayin Zhang: Data curation, Writing – review & editing. Ignacio Despujol: Data curation, Writing – review & editing. Jorge Maldonado-Mahauad: Data curation, Writing – review & editing. German Montoro: Data curation, Writing – review & editing. Melanie Peffer: Data curation, Writing – review & editing. Tobias Rohloff: Data curation, Writing – review & editing. Jenny Lane: Data curation, Writing – review & editing. Carlos Turro: Data curation, Writing – review & editing. Xitong Li: Data curation, Writing – review & editing. Mar Pérez-Sanagustín: Data curation, Writing – original draft, Writing – review & editing. Justin Reich: Conceptualization, Methodology, Investigation, Writing – review & editing, Funding acquisition.

Acknowledgements

We would like to thank support from the MIT-SPAIN program sponsored by "la Caixa" Foundation SEED FUND. José A. Ruipérez-Valiente acknowledges support from the Spanish Ministry of Science and Innovation through the Juan de la Cierva Incorporación program (IJC2020-044852-I). Xitong Li acknowledges funding support from the French National Research Agency (ANR) [Grants ANR AAPG iMOOC-18-CE28-0020-01 and Investissements d'Avenir LabEx Ecodec Grant ANR-11-LABX-0047].

References

Alrajhi, L., Alharbi, K., & Cristea, A. I. (2020). A multidimensional deep learner model of urgent instructor intervention need in mooc forum posts. In *International conference on intelligent tutoring systems* (pp. 226–236). Springer.

Archer, E., & Prinsloo, P. (2020). Speaking the unspoken in learning analytics: Troubling the default. *Assessment & Evaluation in Higher Education, 45*, 888–900.

Aydin, C. H., & Kayabaş, B. K. (2018). Designing culturally sensitive massive open online courses: Learning culture and moocs in Turkey. In *Supporting multiculturalism in open and distance learning spaces* (pp. 208–221). IGI Global.

Bayeck, R. Y., & Choi, J. (2018). The influence of national culture on educational videos: The case of moocs. *International Review of Research in Open and Distributed Learning, 19*.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edx's first mooc. *Research & Practice in Assessment, 8*, 13–25.

Brown, M., McCormack, M., Reeves, J., Brook, D. C., Grajek, S., Alexander, B., Bali, M., Bulger, S., Dark, S., Engelbert, N., et al. (2020). *2020 educause horizon report teaching and learning edition* (Technical Report EDUCAUSE).

Buckingham Shum, S. (2012). *Unesco policy brief: Learning analytics (no. november)*. UNESCO Institute for Information Technologies in Education.

Cagiltay, N. E., Cagiltay, K., & Celik, B. (2020). An analysis of course characteristics, learner characteristics, and certification rates in mitx moocs. *International Review of Research in Open and Distributed Learning, 21*, 121–139.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Miceli Barone, A., & Gialama, M. (2017). A comparative quality evaluation of pbsmt and nmt using professional translators. In *Proceedings of MT summit XVI* (Vol. 1, pp. 116–131). Research Track.

Chuang, I., & Ho, A. (2016). *HarvardX and MITx: Four years of open online courses–fall 2012-summer 2016*. Technical Report. Available at: SSRN: https://ssrn.com/abstract=2889436.

Colas, J.-F., Sloep, P. B., & Garreta-Domingo, M. (2016). The effect of multilingual facilitation on active participation in moocs. *International Review of Research in Open and Distributed Learning, 17*, 280–314.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). Mooc dropout prediction using machine learning techniques: Review and research challenges. In *IEEE global engineering education conference (EDUCON)* (pp. 1007–1014). IEEE.

Dao, X.-Q., Le, N.-B., & Nguyen, T.-M.-T. (2021). Ai-powered moocs: Video lecture generation. In *2021 3rd international conference on image, video and signal processing* (pp. 95–102).

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., Seaton, D. T., & Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*, 56–63.

Davis, D., Chen, G., Jivet, I., Hauff, C., & Houben, G. J. (2016). Encouraging metacognition & self-regulation in moocs through increased learner feedback. In *International conference on learning analytics and knowledge (LAK)* (pp. 17–22).

Dillahunt, T. R., Wang, B. Z., & Teasley, S. (2014). Democratizing higher education: Exploring mooc use among those who cannot afford a formal education. *The International Review of Research in Open and Distributed Learning, 15*.

Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning, 32*, 281–290.

European Commission's Erasmus+ for Higher Education. (2021). *FOSTWOM project. Connecting Women & STEM*. https://fostwom.eu/.

Fabes, R. A., Martin, C. L., Hanish, L. D., Galligan, K., & Pahlke, E. (2015). Gender-segregated schooling: A problem disguised as a solution. *Educational Policy, 29*, 431–447.

Fabes, R. A., Pahlke, E., Martin, C. L., & Hanish, L. D. (2013). Gender-segregated schooling and gender stereotyping. *Educational Studies, 39*, 315–319.

García-Peñalvo, F. J., Fidalgo-Blanco, Á., & Sein-Echaluce, M. L. (2018). An adaptive hybrid mooc model: Disrupting the mooc concept in higher education. *Telematics and Informatics, 35*, 1018–1030.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018). Morf: A framework for predictive modeling and replication at scale with privacy-restricted mooc data. In *2018 IEEE international conference on big data (big data)* (pp. 3235–3244). IEEE.

Godwin-Jones, R. (2014). Global reach and local practice: The promise of moocs. *Language, Learning and Technology, 18*, 5–15.

Gonzalez, C., García-Holgado, A., Plaza, P., Castro, M., Peixoto, A., Merino, J., Sancristobal, E., Menacho, A., Urbano, D., Blazquez, M., et al. (2021). Gender and steam as part of the mooc steam4all. In *2021 IEEE global engineering education conference (EDUCON)* (pp. 1630–1634). IEEE.

Grandon, E. E., Alshare, K., & Kwun, O. (2005). Factors influencing student intention to adopt online classes: A cross-cultural study. *Journal of Computing Sciences in Colleges, 20*, 46–56.

Gupta, K. P. (2019). *Investigating the adoption of moocs in a developing country: Application of technology-user-environment framework and self-determination theory*. Interactive Technology and Smart Education.

Hernández, J., Rodríguez, F., Hilliger, I., & Pérez-Sanagustín, M. (2018). Moocs as a remedial complement: Students' adoption and learning outcomes. *IEEE Transactions on Learning Technologies, 12*, 133–142.

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture, 2*, 2307–0919.

Huang, F., Sánchez-Prieto, J., Teo, T., García-Peñalvo, F. J., Olmos-Migueláñez, S., & Zhao, C. (2021). A cross-cultural study on the influence of cultural values and teacher beliefs on university teachers'information and communications technology acceptance. *Educational Technology Research & Development, 69*, 1271–1297.

Huang, F., Sánchez-Prieto, J., Teo, T., García-Peñalvo, F. J., Sánchez, E. M. T., & Zhao, C. (2020). The influence of university students' learning beliefs on their intentions to use mobile technologies in learning: A study in China and Spain. *Educational Technology Research & Development, 68*, 3547–3565.

Huang, F., Teo, T., Sánchez-Prieto, J. C., García-Peñalvo, F. J., & Olmos-Migueláñez, S. (2019). Cultural values and technology adoption: A model comparison with university teachers from China and Spain. *Computers & Education, 133*, 69–81.

Hunt, A., & Tickner, S. (2015). Cultural dimensions of learning in online teacher education courses. *Journal of Open, Flexible, and Distance Learning, 19*, 25–47.

İnan, E., & Ebner, M. (2020). Learning analytics and moocs. In P. Zaphiris, & A. Ioannou (Eds.), *Learning and collaboration technologies. Designing, developing and deploying learning experiences* (pp. 241–254). Cham: Springer International Publishing.

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate: Preparing learning analytics dashboards for educational practice. In *International conference on learning analytics and knowledge (LAK)* (pp. 31–40).

Kitto, K., Sarathy, N., Gromov, A., Liu, M., Musial, K., & Shum, S. B. (2020). Towards skills-based curriculum analytics: Can we automate the recognition of prior learning?. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 171–180).

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*, Article e0239766.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in moocs. *Science, 355*, 251–252.

Kumar, V. S., & Boulanger, D. (2020). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 1–47.

LeBlanc, Paul (2021). *What 2U's $800 million deal to acquire EdX means for higher*. https://www.forbes.com/sites/paulleblanc/2021/07/06/what-2us-800-million-deal-to-acquire-edx-means-for-higher-ed/.
Liu, Z., Brown, R., Lynch, C., Barnes, T., de Baker, R. S. J., Bergner, Y., & McNamara, D. S. (2016). Mooc learner behaviors by country and culture; an exploratory analysis. In *EDM* (pp. 127–134).
Lohr, S. (2020). *Remember the MOOCs? After near-death, they're booming*. https://www.nytimes.com/2020/05/26/technology/moocs-online-learning.html.
Ma, L., & Lee, C. S. (2019). Understanding the barriers to the use of moocs in a developing country: An innovation resistance perspective. *Journal of Educational Computing Research, 57*, 571–590.
Marsh, H. W., Hau, K.-T., & Kong, C.-K. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39*, 727–763.
Na, K. S., & Tasir, Z. (2017). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In *IEEE conference on big data and analytics (ICBDA)* (pp. 118–123). IEEE.
Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education, 25*, 229–248.
Oh, E. G., Chang, Y., & Park, S. W. (2019). Design review of moocs: Application of e-learning design principles. *Journal of Computing in Higher Education*, 1–21.
Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you* (Penguin UK).
Pérez-Álvarez, R. A., Maldonado-Mahauad, J., Sharma, K., Sapunar-Opazo, D., & Pérez-Sanagustín, M. (2020). Characterizing learners' engagement in moocs: An observational case study using the notemyprogress tool for supporting self-regulation. *IEEE Transactions on Learning Technologies, 13*, 676–688.
Pérez-Martín, J., Rodriguez-Ascaso, A., & Molanes-López, E. M. (2020). Quality of the captions produced by students of an accessibility mooc using a semi-automatic tool. *Universal Access in the Information Society*, 1–14.
Reich, J. (2015). Rebooting mooc research. *Science, 347*, 34–35.
Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science, 363*, 130–131.
Reidmiller, D., Avery, C., Easterling, D., Kunkel, K., Lewis, K., Maycock, T., & Stewart, B. (2017). *Fourth national climate assessment. Volume II: Impacts, Risks, and Adaptation in the United States*. Report-in-Brief.
Rodriguez, B. C. P., Armellini, A., & de la Garza Escamilla, S. L. (2018). Sustainability of massive open online courses (moocs): Beyond business models. In *EdMedia+ innovate learning* (pp. 1641–1647). Association for the Advancement of Computing in Education (AACE).
Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*, 1–21.
Rosen, Y., Rushkin, I., Rubin, R., Munson, L., Ang, A., Weber, G., Lopez, G., & Tingley, D. (2018). Adaptive learning open source initiative for mooc experimentation. In *International conference on artificial intelligence in education* (pp. 307–311). Springer.
Ruipérez-Valiente, J. A. (2021). *Supplementary materials for paper: Global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions*. https://doi.org/10.17605/OSF.IO/MBWDX. https://osf.io/mbwdx/
Ruiperez-Valiente, J. A., Halawa, S., & Reich, J. (2019). Multiplatform MOOC analytics: Comparing global and regional patterns in edX and Edraak. In *Proceedings of the sixth ACM conference on Learning@Scale*. ACM.
Ruipérez-Valiente, J. A., Halawa, S., Slama, R., & Reich, J. (2020a). Using multi-platform learning analytics to compare regional and global mooc learning in the arab world. *Computers & Education, 146*, 103776.
Ruipérez-Valiente, J. A., Jenner, M., Staubitz, T., Li, X., Rohloff, T., Halawa, S., Turro, C., Cheng, Y., Zhang, J., Despujol, I., et al. (2020b). Macro MOOC learning analytics: Exploring trends across global and regional providers. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 518–523).
Santos Espino, J. M., Guerra Artal, C., & González Betancor, S. M. (2021). Video lectures: An analysis of their useful life span and sustainable production. *International Review of Research in Open and Distributed Learning, 22*, 99–118.
Saville-Troike, M. (2003). Extending "communicative" concepts in the second language curriculum, a sociolinguistic perspective. In D. Lange (Ed.), *Culture as the core: Perspective on Culture in second language education research in second language learning*. Information Age Publishing (Incorporated).
Shah, D., & Pickard, L. (2019). *Massive list of MOOC providers around the world*. https://www.classcentral.com/report/mooc-providers-list/.
Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*, 1380–1400.
Society of Learning Analytics Research (SoLAR). (2021). *What is learning analytics?*. www.solaresearch.org/about/what-is-learning-analytics/.
Sosoni, V., & Stasimioti, M. (2016). *Translation as a microtask: Investigating a paradox*. https://www.academia.edu/29700551/Translation_as_a_Microtask_Investigating_a_Paradox. Online. (Accessed 31 January 2021).
St Amant, K. (2007). Online education in an age of globalization: Foundational perspectives and practices for technical communication instructors and trainers. *Technical Communication Quarterly, 16*, 13–30.
Steffens, K. (2015). Competences, learning theories and moocs: Recent developments in lifelong learning. *European Journal of Education, 50*, 41–59.
Sulkowski, N., & Deakin, M. K. (2009). Does understanding culture help enhance students' learning experience? *International Journal of Contemporary Hospitality Management, 21*, 154–166.
The Advanced Distributed Learning Initiative. (2013). *Experience api standard overview*. https://adlnet.gov/research/performance-tracking-analysis/experience-api/.
Tremblay, A., Broersma, M., Coughlin, C. E., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology, 7*, 985.
Uchidiuno, J., Koedinger, K., Hammer, J., Yarzebinski, E., & Ogan, A. (2018). How do English language learners interact with different content types in mooc videos? *International Journal of Artificial Intelligence in Education, 28*, 508–527.
Uchidiuno, J. O., Ogan, A., Yarzebinski, E., & Hammer, J. (2018). Going global: Understanding English language learners' student motivation in English-language moocs. *International Journal of Artificial Intelligence in Education, 28*, 528–552.
Veletsianos, G., & Shepherdson, P. (2016). A systematic analysis and synthesis of the empirical mooc literature published in 2013–2015. *The International Review of Research in Open and Distributed Learning, 17*.
Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*, 910–932.
Wong, J.-S., et al. (2018). Messagelens: A visual analytics system to support multifaceted exploration of mooc forum discussions. *Visual Informatics, 2*, 37–49.
Xing, W., & Du, D. (2019). Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*, 547–570.
Yuan, L., & Powell, S. (2013). *MOOCs and open education: Implications for higher education*. http://publications.cetis.ac.uk/2013/667.
van der Zee, T., & Reich, J. (2018). Open education science. *AERA Open, 4*, 2332858418787466.
Zhang, K., Bonk, C. J., Reeves, T. C., & Reynolds, T. H. (2019). *MOOCs and open education in the global South: Challenges, successes, and opportunities*. Routledge.