# *Macro MOOC learning analytics: exploring trends across global and regional providers*

# Macro MOOC Learning Analytics: Exploring Trends Across Global and Regional Providers

José A. Ruipérez-Valiente
University of Murcia and MIT

Matt Jenner
FutureLearn

Thomas Staubitz
Hasso Platter Institute

Xitong Li
HEC Paris

Tobias Rohloff
Hasso Platter Institute

Sherif Halawa
Edraak

Carlos Turro
Polithecnic University of Valencia

Yuan Cheng
Tsinghua University

Jiayin Zhang
Tsinghua University

Ignacio Despujol
Polithecnic University of Valencia

Justin Reich
MIT

## ABSTRACT

Massive Open Online Courses (MOOCs) have opened new educational possibilities for learners around the world. Numerous providers have emerged, which usually have different targets (geographical, topics or language), but most of the research and spotlight has been concentrated on the global providers and studies with limited generalizability. In this work we apply a multi-platform approach generating a joint and comparable analysis with data from millions of learners and more than ten MOOC providers that have partnered to conduct this study. This allows us to generate learning analytics trends at a macro level across various MOOC providers towards understanding which MOOC trends are globally universal and which of them are context-dependent. The analysis reports preliminary results on the differences and similarities of trends based on the country of origin, level of education, gender and age of their learners across global and regional MOOC providers. This study exemplifies the potential of macro learning analytics in MOOCs to understand the ecosystem and inform the whole community, while calling for more large scale studies in learning analytics through partnerships among researchers and institutions.

## CCS CONCEPTS

• **Applied computing** → **Distance learning**; **E-learning**; • **Information systems** → *Data mining*; • **Social and professional topics** → *User characteristics.*

## KEYWORDS

MOOCs; Learning Analytics; Multi-plaform Analytics Collaboration; Large-scale Analytics; Cultural Factors



**ACM Reference Format:**
José A. Ruipérez-Valiente, Matt Jenner, Thomas Staubitz, Xitong Li, Tobias Rohloff, Sherif Halawa, Carlos Turro, Yuan Cheng, Jiayin Zhang, Ignacio Despujol, and Justin Reich. 2020. Macro MOOC Learning Analytics: Exploring Trends Across Global and Regional Providers. In *Learning Analytics & Knowledge Conference 2020, Frankfurt (Germany).*

## 1 INTRODUCTION

The rise of MOOCs has widened the educational landscape with new opportunities. One original promise of MOOCs was to provide high quality, free educational resources around the world, especially to those learners lacking ready, affordable access to higher education [6]. However, many studies have reported that most MOOC learners are already educated and from affluent countries [11]. Most of these studies have focused on global MOOC providers (such as edX, FutureLearn or Coursera), where Anglo-American higher education universities teach courses primarily in English. However, very few studies have delved into differences with local or regional MOOC providers, that center their attention on a local or regional population. Many institutions and national MOOC initiatives are using the open source software Open edX, which provides an easy way to publish their courses. As of 2018, there are over 800 organizations, institutions, and governments which have been running instances of Open edX [3]. There are numerous studies that have discussed the impact of language and culture in learning [8], and previous researchers have linked the country of origin of MOOC participants to different behavioral patterns in the course [10] or to social identity threat in less developed countries [9]. Previous work that compared Arab learners in both Edraak (an Arabic MOOC provider) and edX found that learner populations in Edraak had a wider range of education levels and a more even gender, and the courses showed more favorable completion trends [12]. This previous work suggested that regional MOOC providers might be better positioned to fulfil their learners' needs as they offer courses in their local language, and taught by instructors of similar culture and background. It may be that regional providers are better positioned to fulfill the democratizing promise of MOOCs than large elite institutions, but research about demographics, readiness, participation, and learning in regional MOOC providers is nascent.

In this paper, we address this challenge through a multi-platform analysis approach with a variety of global and regional MOOC providers.

Buckingham Shum introduced three levels where learning analytics can have an impact, the macro, meso and micro [2]. Additionally, Drachsler and Kalz mapped those levels to the MOOC and Learning Analytics Cycle (MOLAC) where the micro level focuses on a single course, the meso a set of MOOCs, and the macro level provides analytics that informs the whole community [7]. With these distinction in mind, we situate our study at a macro level of MOOC learning analytics, providing high level demographic trends for more than ten MOOC providers, generating a study with insights that can inform the whole community. Prior studies in MOOC research often focused on a detailed analysis of one or a few courses (e.g. [1]), which do not allow to generalize, longitudinal studies with many courses from one single MOOC provider (e.g. [4, 5], which do not capture differences among MOOC providers or literature reviews of MOOC analytic studies [13], which are not comparable as different methods are applied in different studies. We believe that one of the most underexplored areas of learning analytics, is understanding variations in trends across virtual learning environments, which should be pushed forward in the coming years. In this study we describe the methodology "Multiplatform MOOC Analytics" that we have applied to put together data and analysis of more than ten MOOC providers. While a simplified version of this method was previously proposed [12], it is the first time that we apply it with a large number of providers and institutions. We also provide preliminary results on how a number of demographic variables are distributed across all of these platforms. Our overarching research question explores the extent to which MOOC trends are globally universal versus context-dependent, and more specifically, we look for differences between global and regional MOOC providers.

## 2 METHODOLOGY

### 2.1 Multiplatform MOOC Analytics

We describe the process that was followed to conduct this research now. First, the project lead launched an initial call looking for partners with access to large MOOC datasets from different platforms with the objective of running a comparative study on global and regional trends. Once the partnership was settled, we followed the next steps to conduct the research:

(1) Partners shape their data into the same common format.
(2) The project lead generates a Jupyter notebook that is expecting the common data format established in the previous step. This script outputs aggregate data from different institutions that is merged together for the joint analysis.
(3) We conduct the joint data analysis of all providers together and iterate over these three steps as required.

The initial call for partners, common data format and additional methodological description can be consulted online[1]. This methodology greatly alleviates the logistical and privacy concerns of sharing student-level information. Additionally, we are able to perform an "apples-to-apples" comparison as datasets contain the same variables and the analysis is conducted using exactly the same script.

[1]Blinded for review

### 2.2 Context and Data Collection

We provide a brief description of the context and the size of data collected of the providers that have joined this partnership thus far:

- **MITx and HarvardX** (abbreviated as MITxHx): Hosted in edX, which was also founded by MIT and Harvard, teaching their courses to a global audience in English. The nature and target of the courses are diverse, with courses in STEM, but also many in the area of social sciences and humanities. Data collected of around 3.7 million learners and 552 MOOCs.
- **FutureLearn**: Founded by the UK Open University and partners with over 170 organisations globally to provide MOOCs, microcredentials and degrees. Most courses are in English. Data collected of around 1.1 million learners and 1548 MOOCs.
- **openHPI**: Since 2012 the platform offers courses about digital technologies, transformation and engineering in German and English as one of the MOOC pioneers in Europe. Based on the HPI MOOC Platform. Data collected of around 113 thousand learners and 43 MOOCs.
- **openSAP**: In 2013 the German-based software company SAP launched their platform for enterprise MOOCs. The primary objective is to enlarge the SAP ecosystem by offering education and for their employees, clients and partners about their products. The majority of courses are offered in English. Based on the HPI MOOC Platform. Data collected of around 515 thousand learners and 166 MOOCs.
- **OpenWHO**: Developed in 2016 in a cooperation between the World Health Organization (WHO) and the Hasso Plattner Institute (HPI). The platform aims to improve the response to health emergencies by providing courses for frontline responders to better contain disease outbreaks. Therefore, courses are offered in a variety of languages. Based on the HPI MOOC Platform. Data collected of around 35 thousand learners and 52 MOOCs.
- **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. Some of the courses are open to the general public learners, while others are required for courses credits. As a partner with Coursera, HEC Paris offers its online courses hosted on the Coursera platform. Data collected of around 22 thousand learners and 33 MOOCs.
- **UPValenciaX**: Supported by Universitat Politecnica de Valencia in Spain and hosted in edX, provides a variety of courses in STEM, nearly all in Spanish. Data collected of around 700 thousand learners and 230 MOOCs.
- **UPVx**: Another site supported by Universitat Politecnica de Valencia which is hosted in its own Open edX instance. Focuses in local topics for the Valencian region and basic STEM courses. Courses are in Catalan (Valencian) and Spanish. Data collected of around 40 thousand learners and 132 MOOCs.
- **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. Data collected of around 24 thousand learners and 18 MOOCs.
- **Edraak**: Edraak was founded in 2013 by the Queen Rania Foundation for Education and Development to surpass the barriers

of learning in English. Edraak produces all of its courses in Arabic, and hosts them on its locally-adapted Open edX platform. Edraak's courses span multiple categories, including STEM, business and workforce development skills, health, arts, and language. Course content is designed in collaboration with regional experts from academia and industry. Data collected of around 610 thousand learners and 228 MOOCs.

- **XuetangX**: XuetangX is the world's first Chinese MOOC platform. Founded by Tsinghua University in 2014, it is authorized to operate edX courses in the Chinese mainland. XuetangX offers courses provided by prestigious Chinese schools, leading universities and institutions abroad. The 2300 courses it offered cover almost all subjects and are mainly in Chinese and English. Data collected of around 655 thousand learners and 2884 MOOCs.
- **The ChineseMOOC**: The Chinese MOOC was launched by a joint effort of Peking University and Alibaba Group in 2015. It was hosted on Alibaba Cloud platform. The online courses are mostly offered in Chinese with the intention to attract Chinese learners from all over the world. Data collected of around 7 thousand learners and two MOOCs.

While the common data format includes more variables, in this preliminary analysis we use the country of origin, age, level of education and gender of learners. Finally, we note that depending on the data each provider captures and if the learners reported these demographics, not all variables are available for all learners.

# 3 RESULTS

## 3.1 Country Representation by Provider

We present here how the country of origin of learners is distributed by provider. Figure 1 shows a stacked bar chart with the top-ten most representative countries in percentage of learners for each platform. Additionally, the color codifies the region of the country, which helps to perceive the regional focus on each provider. Several key trends emerge: we find that both MITxHx and FutureLearn have similar baseline particularly from their home countries, about 30% of learners. We see a certain level of similarity in the most representative countries in all global MOOC platforms, with USA, UK, India or Brazil being in the top of all of them. Perhaps the exception is OpenWHO, where maybe the nature of the courses focusing on world health issues attracts a more diverse population from different regions.

Those providers that share courses in both English and a local language, have predominantly learners from the local region, but also from other regions. See HEC Paris with French population, openHPI and mooc.house with the German, and UPValenciaX with Hispanic population. An interesting follow-up for UPValenciaX and UPVx is that, although they have very similar courses in nature, UPValenciaX that is hosted on edX has a much more international audience from many Hispanic countries when compared to UPVx, which is a more local initiative of the university and has predominantly Spanish learners. On other hand, we see that the providers that focus only on a specific region, like Edraak, XuetangX and the ChineseMOOC, primary bring learners from those regions. In the case of Edraak, all countries are within the Arab region, and for XuetangX learners are primarily based in China. In the case of the ChineseMOOC, the population mainly comes from China, but also from diverse countries in Asia and USA, perhaps because the ChineseMOOC seeks to achieve Chinese learners from all over the world. These distributions demonstrate that the different global and regional providers have distinct missions and use diverse strategies to recruit students from different geographic regions.

## 3.2 Level of Education by Region and Provider

In the next figure 2 we show the distribution of the level of education in a 100% stacked bar chart. Due to the differences between educational systems, some less established educational categories were not comparable across providers (such as specializations or associate degrees), thus we remove them in order to focus on those that we can compare and are well-established across all educational systems. We present four different educational levels, 'Doctorate', 'Master', 'Bachelor' and 'High school, junior high school or elementary school (HS/JHS/EL)', that we represent in a green divergent palette of colors (darker means higher level of education).

An overall trend that has been reported in several studies is that Europe and Northern America learners have higher levels of education at a doctorate or master level [4], which we can see that is quite constant across all MOOC providers. There are interesting distinctions when comparing global providers. `MITxHx` and FutureLearn show similar proportions of learners with a doctorate or master, but `MITxHx` attracts more learners with only an HS/JHS/EL education, when compared with FutureLearn. Additionally, openSAP has less learners with a doctorate or HS/JHS/EL education, and most of them have a bachelor or master level.

The regional providers Edraak and XuetangX have the widest range of education levels and most learners with lower levels of education, with 86% and 79% of their learners respectively with a bachelor or HS/JHS/EL education. Also is interesting to see how the European population of openHPI has a bimodal distribution with highly educated learners with a doctorate or a master on one side, and HS/JHS/EL learners on the other side. UPVx shows a clear difference between the more educated learners from Spain and the Spanish speakers from Latin America. Another interesting difference is why UPVx attracts more educated learners from both Spain and Latin America than UPValenciaX. These demographic observations are in agreement with some trends reported previously in the literature, and open new questions about potential causes of variation.

## 3.3 Gender by Region and Provider

Figure 3 shows the distribution of gender by region and provider in a 100% stacked bar chart using two colors. Another one of the trends that the literature has reported with frequency, is more acute gender gaps, with a higher proportion of male learners in regions with lower levels of human development [4]. Then, regions like Europe or Northern America often have a better gender balance than regions like Africa or the Arab countries. We see that this pattern is consistent for some of the providers like `MITxHx`, FutureLearn, UPValenciaX or HEC Paris. However, in the case of openSAP or openHPI we see that the gender gap is systematically low for all regions, while in the case of OpenWHO we see how Arabic and Latin America regions have higher female representation than European or Northern American regions. We believe that these can be

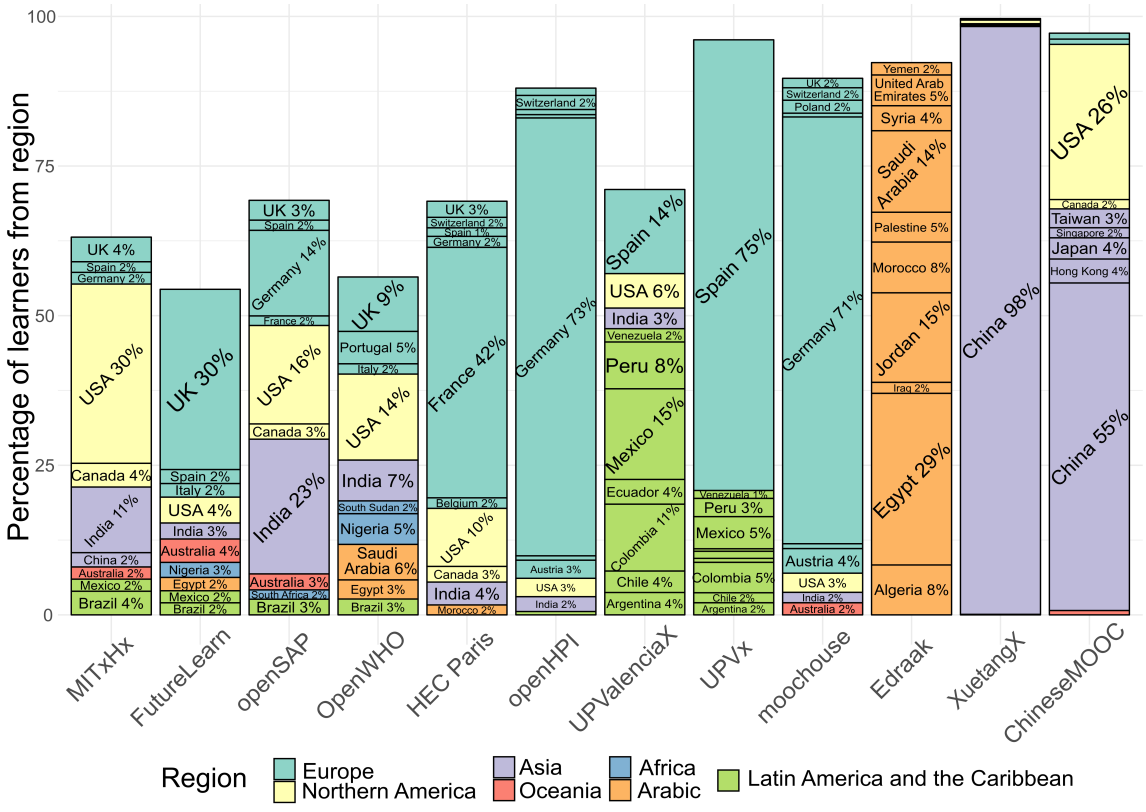


**Figure 1: Top-ten most representative countries in percentage per provider. Color codifies the region of the country.**

influenced by the nature of the courses, with openSAP and openHPI being very focused on technical courses, while OpenWHO provides courses on world health issues, and hence these can attract systematically different demographics of learners than other platforms. Delving into the factors that are affecting these gender distribution differences across providers can help in designing learning experiences that reduce the current gender gaps.

## 3.4 Age Range by Provider

We also explore the distribution of age by provider in Figure 4 in a 100% stacked bar chart. We cluster learners in different age buckets and codify those buckets in a blue divergent palette of colors (darker means older) and thus the comparison across providers is straightforward. The most common age bucket for most providers is [26, 35), except for openHPI with [45, 55) and Edraak with [18, 25). The trend shows that the regional MOOC providers Edraak, XuetangX and UPValenciaX, together with the global `MITxHx` have the youngest populations of learners. On the other side, providers HEC Paris, openHPI, mooc.house and the global FutureLearn, have the oldest population of learners. Additionally, FutureLearn shows the most heterogeneous distribution of learners in terms of age. The rest of providers openSAP, OpenWHO and UPVx have mainly young professionals within the age interval of [26, 45). While some of these differences might be related to the age target of providers and their courses, regional variations can also be linked to digital literacy and level of English knowledge across ages.

## 4 DISCUSSION AND CONCLUSIONS

This multiplatform analysis represents an important early step in the global analysis of the MOOC phenomenon through large-scale, cross-provider data analysis of MOOCs. The investments made into these platforms, the courses and each learner are substantial, so learning analytics researchers should continue to advance methods and approaches that enhance our understanding of the overall ecosystem. By collaborating on this global multiplatform research study we provide a view into MOOCs that was otherwise challenging or unavailable. Our main findings suggest that age, gender, level of education and region can, in aggregate, provide vital information about the types of learners taking MOOCs and the value of providers across local and global populations. However, the aim of this research was to unlock the value of comparison between providers and gain new insights into global and local learners alike. Benchmarking is likely a useful outcome from this research – where providers and course teams can compare their demographics against these published datasets with the normalised set of figures. However, it is anticipated that this research will start to unlock new understanding on regional and global online learning.

We can see from this study there is an impact of the locality of the platform. Platforms have very different catchment areas for their courses, with varying levels of concentration. This concentration of home country participation ranges from as high as 98% for XuetangX, to 30% for the global providers. Exploring the reasons for this would help understand when providers differ between a

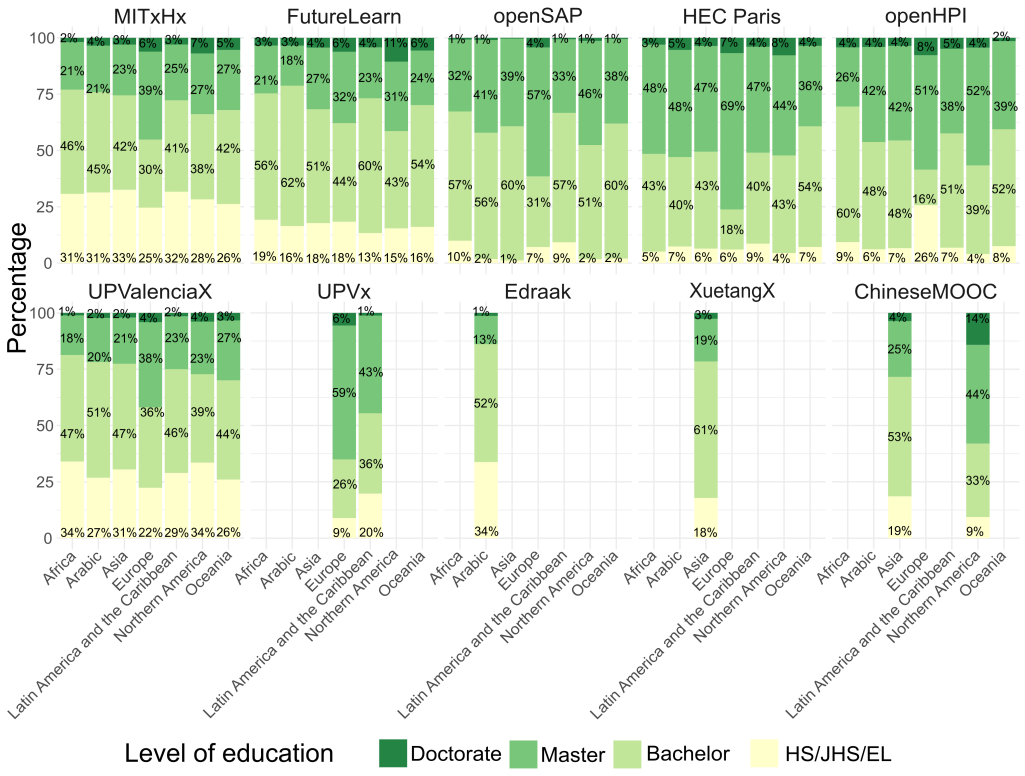


**Figure 2: Distribution of level of education in percentage per provider and region.**

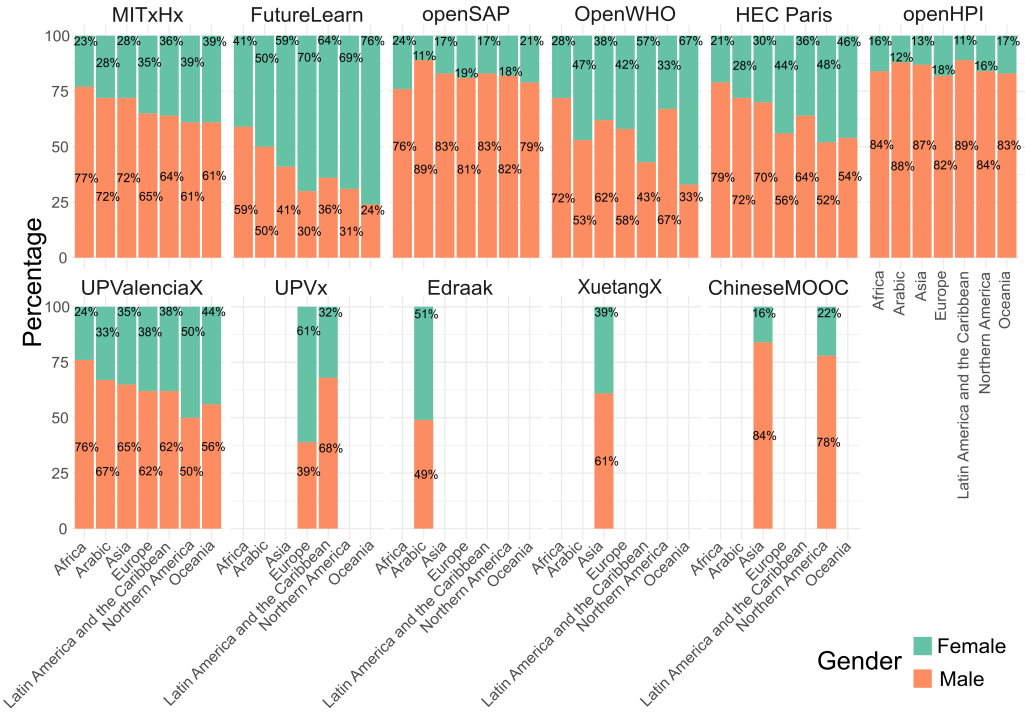


**Figure 3: Distribution in percentage of gender per provider and region.**

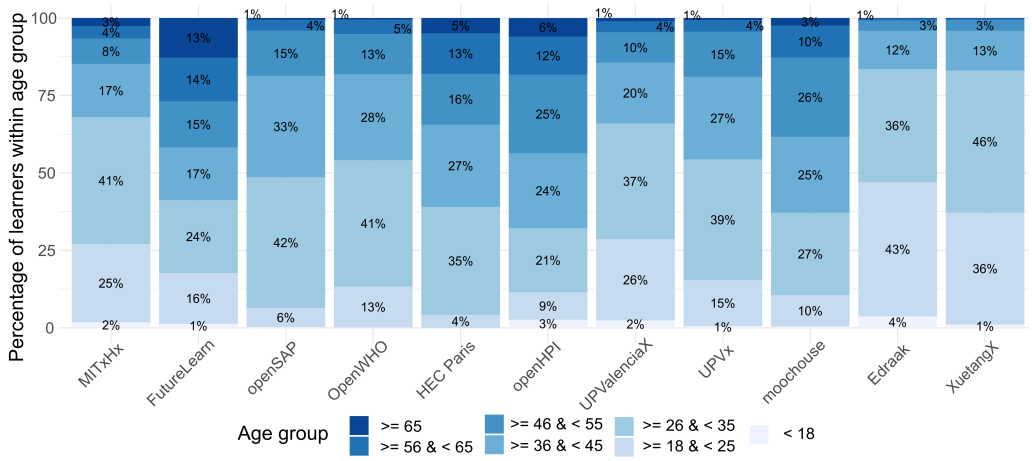


**Figure 4: Distribution of learners in percentage within each age group by provider.**

local or global focus or either want to shift from being local to becoming more international or to hone in on a specific region or demographic in reach or appeal. Gender balance is one indicator of how each platform has managed to attract different audiences. Overall, participation by gender is imbalanced, with average of 63% of learners identifying as male across all platforms. On some platforms, notably openSAP, openHPI and ChineseMOOC, this imbalance is significantly larger with an average 79% male learners. FutureLearn ranks as the platform with the largest percentage of learners who identify as female and OpenWHO, UPVx and Edraak have notably better gender balanced demographics. The analysis does not intentionally exclude learners who identify as fluid or non-binary – these data are not available in sufficient quantities from enough providers. Further studies would benefit from providers being able to collect a wider dataset in this area and for caution in analysis when using traditional binary classification. The difference of level of previous education across learners proves that MOOC providers can generate interest across a wide audience of diverse prior education levels. While there is value in understanding goals and motivations of well educated individuals conducting lifelong learning in MOOCs, we should aim to understand how some of these providers are reaching less educated learners that might not otherwise have access to high quality education to learn from these best practices – knowing which platforms recruit the widest range of learners is one important descriptive step towards understanding these best practices. All of these are elements related to the global issue of designing more equitable and inclusive online learning experiences.

A number of factors might be affecting these demographic differences across MOOC providers, such as the concentration of certain topics in the course catalog, instructional design, language of instruction or geographical location. We know each learner has their own motivations and goals for taking a MOOC, yet in aggregate we can also look for patterns to learn from as education researchers, especially when using a common dataset with millions of records distributed across platforms. More research is needed in macro learning analytics and regional MOOC providers to fully appreciate the influence these factors are making in the learners that register to these courses and the quality of their learning process. By understanding learners at the macro level, it may be possible to further increase learning outcomes and performance for MOOC providers at platform and individual levels too.

This study used a set of common metrics between different platforms. To expand on this work we had to ensure we could understand, and accurately analyse the differences in how each platform collects key operational metrics. An enrolment, for example, may be similar across all platforms – but other metrics such as completion, viewing, active learning and grades can be understood in different ways. The authors are working on a legend / key that will enable multi-platform MOOC analysis for our further research and publication for other researchers too. Additional future steps include linking these headline demographics datasets to a deeper exploration of in-course behaviours and processes. This will initially include alignment to the activation, progress and completion each individual learner makes when taking courses with the providers in this study. It is anticipated the further research will unearth local and global patterns in how learners learn and explore what factors may lead to higher levels of interaction and engagement. Despite these results are at a preliminary stage, we share our enthusiasm towards the potential of conducting learning analytics at a macro scale, while encouraging the community to perform more large scale studies through partnerships between researchers and institutions to advance the field forward.

## REFERENCES

[1] Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, and Andrew D Ho. 2013. Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment* 8 (2013), 13–25.

[2] SB Buckingham Shum. 2012. UNESCO Policy Brief: Learning Analytics (No. November). *UNESCO Institute for Information Technologies in Education. Retrieved from www. iite. unesco. org/publications/3214711* (2012).

[3] Class Central. 2018. Sessions We're Excited About at the 2018 Open edX Conference. https://www.classcentral.com/report/2018-open-edx-conference

[4] Isaac Chuang and Andrew Ho. 2016. HarvardX and MITx: Four years of open online courses–fall 2012-summer 2016. (2016).

[5] Alexandra I Cristea, Ahmed Alamri, Mizue Kayama, Craig Stewart, Mohammad Alshehri, and Lei Shi. 2018. Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses. In *27th International Conference on Information Systems Development (ISD2018)*. Springer.

[6] Tawanna R Dillahunt, Brian Zengguang Wang, and Stephanie Teasley. 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, 5 (2014).
[7] Hendrik Drachsler and Marco Kalz. 2016. The MOOC and learning analytics innovation cycle (MOLAC): a reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning* 32, 3 (2016), 281–290.
[8] Anthony Hunt and Sue Tickner. 2015. Cultural dimensions of learning in online teacher education courses. *Journal of Open, Flexible, and Distance Learning* 19, 2 (2015), 25–47.
[9] René F Kizilcec, Andrew J Saltarelli, Justin Reich, and Geoffrey L Cohen. 2017. Closing global achievement gaps in MOOCs. *Science* 355, 6322 (2017), 251–252.
[10] Zhongxiu Liu, Rebecca Brown, Collin Lynch, Tiffany Barnes, Ryan Shaun Joazeiro de Baker, Yoav Bergner, and Danielle S. McNamara. 2016. MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis. In *EDM*. 127–134.
[11] Justin Reich and José A Ruipérez-Valiente. 2019. The MOOC Pivot. *Science* 363, 6423 (2019), 130–131.
[12] Jose A Ruiperez-Valiente, Sherif Halawa, and Justin Reich. 2019. Multiplatform MOOC Analytics: Comparing Global and Regional Patterns in edX and Edraak. In *Proceedings of the Sixth ACM Conference on Learning@Scale*. ACM.
[13] George Veletsianos and Peter Shepherdson. 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, 2 (2016).