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 Plattner Institute

Xitong Li
HEC Paris

Tobias Rohloff
Hasso Plattner 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. Most of the research and spotlight has been concentrated on a handful of global, English-language providers, but there are a growing number of regional providers of MOOCS in languages other than English. In this work, we have partnered with thirteen MOOC providers from around the world. We apply a multi-platform approach generating a joint and comparable analysis with data from millions of learners. This allows us to examine learning analytics trends at a macro level across various MOOC providers, with a goal of 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

LAK'20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/3375462.3375482

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 *LAK'20: Learning Analytics and Knowledge Conference, March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375462.3375482

1 INTRODUCTION

The rise of massive open online courses (MOOCs) has created new opportunities in the educational landscape. One of the original promises of MOOCs was to provide high quality, free educational resources around the world, especially to those learners lacking ready, affordable access to higher education [4]. However, many studies have reported that most MOOC learners already have higher education credentials and come from affluent countries [9]. 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. There are numerous studies that have discussed the impact of language and culture in learning [6], and previous researchers have linked the country of origin of MOOC participants to different behavioral patterns in the course [8] or to social identity threat in less developed countries [7]. 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 ratio, and the courses showed more favorable completion trends [11]. 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, taught by instructors of similar culture and background [11]. 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, by combined data from 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 [5]. 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 have often focused on a detailed analysis of one or a few courses (e.g. [1]) or longitudinal studies with many courses from one single MOOC provider (e.g. [3]. These have limited generalizability to the entire field of open online learning, and they do not capture differences among MOOC providers. There are some literature reviews that have attempted to provide a unified look at MOOC analytic studies [12], but there are limits to comparisons conducted with different methods as applied in different studies. An underexplored area of learning analytics is understanding variations in trends across virtual learning environments. In this study we describe a methodology for "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 [10], this paper presents the first results from a large number of MOOC providers and institutions. We also provide preliminary results on how several 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

In this section, we describe the process that was followed to conduct this research. 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 shaped their data into the same common format.
(2) The project lead generated a Jupyter notebook that takes as input the common data format established in the previous step.
(3) This script outputs aggregate data from different institutions that was merged together for the joint analysis.
(4) We conducted the joint data analysis of all providers together and iterated over these four steps as required.

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 our datasets contain the same variables and our analysis is conducted using exactly the same script.

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): The two original partners in the edX consortium. The majority of courses are taught to to a global audience in English. Study data includes 3.7 million learners from 552 MOOC instances.
- **FutureLearn**: Founded by the UK Open University, with over 170 partner institutions globally to provide MOOCs, microcredentials and degrees. Most courses are in English. Study data includes 1.1 million learners from 1548 MOOC instances.
- **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. Study data includes 113 thousand learners from 43 MOOC instances.
- **openSAP**: In 2013 the German-based software company SAP launched their platform for enterprise MOOCs. Based on the HPI MOOC Platform. Study data includes 515 thousand learners from 166 MOOC instances.
- **OpenWHO**: Developed in 2016 in a cooperation between the World Health Organization (WHO) and the HPI. The platform aims to improve the response to health emergencies with courses offered in a variety of languages. Based on the HPI MOOC Platform. Study data includes 35 thousand learners from 52 MOOC instances.
- **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. Study data includes 24 thousand learners from 18 MOOC instances.
- **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. Study data includes 22 thousand learners from 33 MOOC instances.
- **UPValenciaX**: Supported by Universitat Politecnica de Valencia in Spain and hosted by edX, provides a variety of courses in STEM, nearly all in Spanish. Study data includes 700 thousand learners from 230 MOOC instances.
- **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. Study data includes 40 thousand learners from 132 MOOC instances.
- **Edraak**: Edraak 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 on its locally-adapted Open edX platform. Study data includes 610 thousand learners from 228 MOOC instances.
- **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. Study data includes 655 thousand learners from 2884 MOOC instances.
- **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. Study data includes 7 thousand learners from two MOOC instances.

While the common data format includes more variables, in this preliminary analysis we use student's country of origin, age, level of education and gender of learners. Not every provider captures all of these fields, and not all learners report all of the requested data, so 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 terms of percentage of learners for each platform. Additionally, the color signifies the region of the country, which helps to identify the regional focus on each provider. Several key trends emerge: we find that both MITxHx and FutureLearn have similar participation from their home countries, about 30% of learners. The world's largest countries are, not surprisingly, the most represented across all global MOOC platforms, with USA, UK, India or Brazil being in the top of all of them. Perhaps the exception is OpenWHO, where the nature of courses focusing on world health issues attracts a more diverse population from different regions.

For the providers that offer courses in both English and a local language, we find that these platforms predominantly have learners from the local region, but also from other regions. Therefore, the HEC Paris mainly has students from the French population, openHPI and mooc.house serve mostly students in Germany, and UPValenciaX serves primarily a 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 markets to 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 are well-established and comparable 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 palette of colors where darker shades represent 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 [3], and we find that this is 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 fewer learners with a doctorate or HS/JHS/EL education, and most of them have a bachelors or masters degree.

The regional providers Edraak and XuetangX have the widest range of education levels and they have the most learners with lower levels of education, with 86% and 79% of their learners respectively with a bachelor or HS/JHS/EL education. Also of interest is how the 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 that UPVx attracts more educated learners from both Spain and Latin America than UPValenciaX. These demographic observations are aligned 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. MOOC gender gaps have been reported frequently in the literature, with a higher proportion of male learners, especially in regions with lower levels of human development [3]. In the global, elite providers, 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 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.

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.

3.4 Age 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 segments and codify those segments in a blue divergent palette of

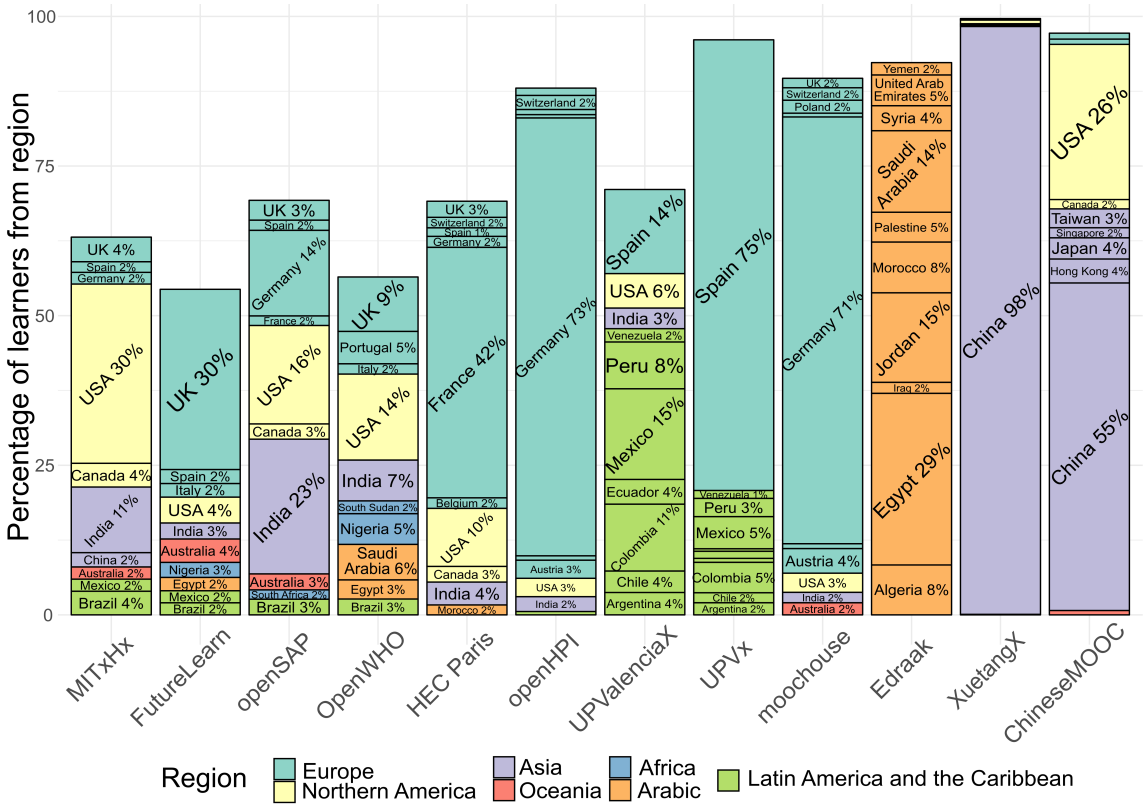

Figure 1: Top-ten most representative countries in percentage per provider. Color codifies the region of the country.

colors (darker means older) allowing a straightforward comparison across providers. The most common age segment 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 might 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 and courses 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 new view into this global ecosystem of MOOC providers. Our main findings suggest that age, gender, level of education and region can, in aggregate, provide useful information about the types of learners taking MOOCs and the value of providers across local and global populations. The primary aim of this research was to unlock the value of comparison between providers and gain new insights into learners from both global and regional providers. This research provides a set of benchmarks for future studies, where providers and course teams can compare their demographics against these published datasets. Ideally, this approach to global research will start to unlock new understanding on regional and global online learning.

We show that locality impacts platforms. 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 these difference would help understand when providers differ between a 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 difference of level of previous education

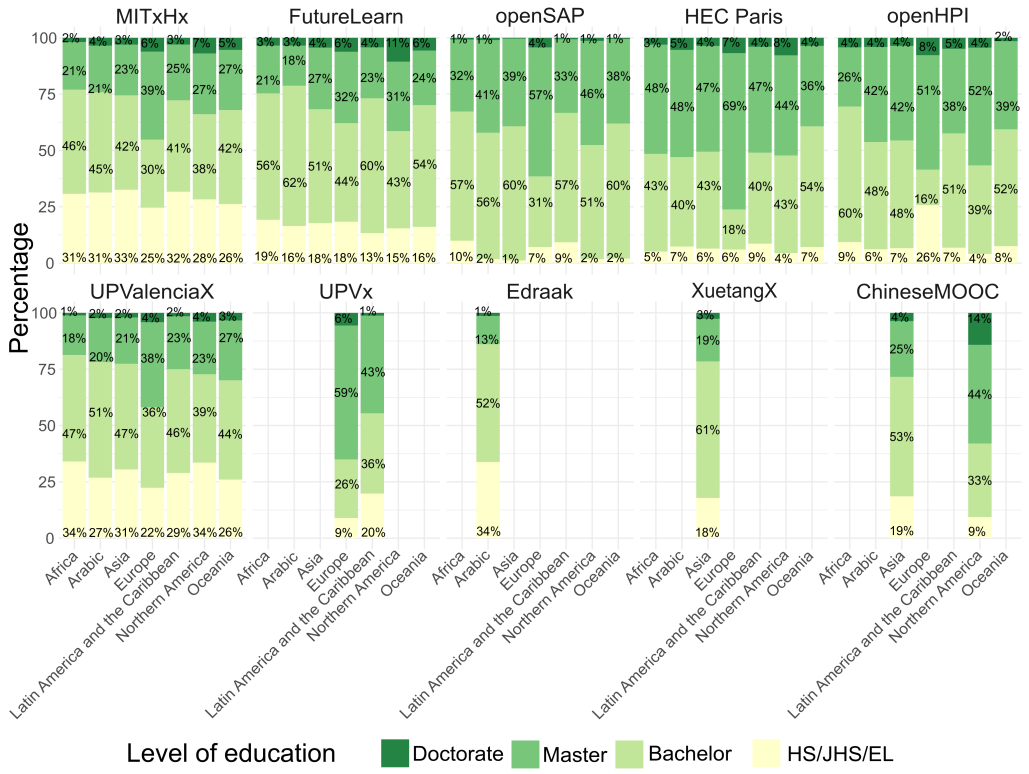

Figure 2: Distribution of level of education in percentage per provider and region.

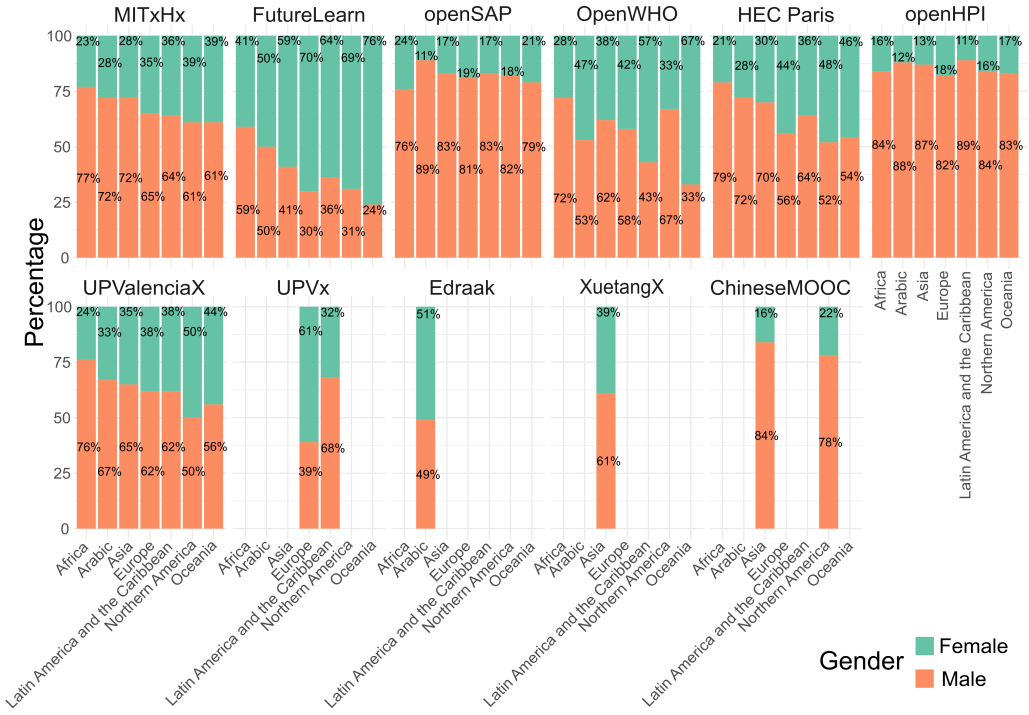

Figure 3: Distribution in percentage of gender per provider and region.

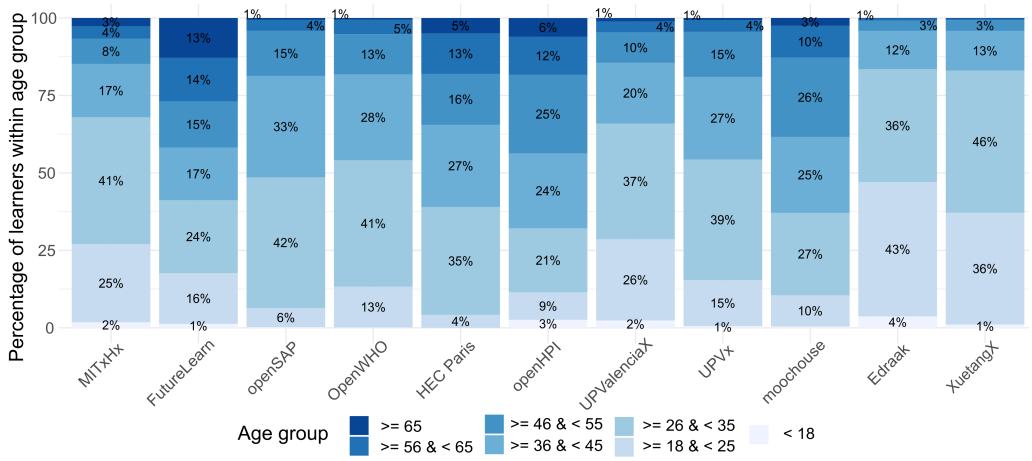

Figure 4: Distribution of learners in percentage within each age group by provider.

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. The goal would be to learn from the providers who recruit the widest range of learners as one important step towards understanding best practices in widening access to education. 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 the demographic differences we find 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 from 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. 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 that 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. These results are at a preliminary stage, but we share them with an enthusiasm for 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.

ACKNOWLEDGMENTS

We would like to thank support from the MIT-SPAIN "la Caixa" Foundation SEED FUND. The first author acknowledges support from the Spanish Ministry of Economy and Competitiveness through the Juan de la Cierva Formación program (FJCI-2017-34926).

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

[4] 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).

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Justin Reich and José A Ruipérez-Valiente. 2019. The MOOC Pivot. *Science* 363, 6423 (2019), 130–131.

[10] José A. Ruipérez-Valiente, Sherif Halawa, and Justin Reich. 2019. Multiplatform MOOC Analytics: Comparing Global and Regional Patterns in edX and Edraak. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S'19)*. Article 3, 9 pages.

[11] José A Ruipérez-Valiente, Sherif Halawa, Rachel Slama, and Justin Reich. 2019. Using multi-platform learning analytics to compare regional and global MOOC learning in the Arab world. *Computers & Education* (2019), 103776.

[12] 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).