Exploring Cross-Country Prediction Model Generalizability in MOOCs

Juan-Miguel Andres-Bray
University of Pennsylvania
Philadelphia PA USA
Miggy.Andres-Bray@mheducation.com

Stephen Hutt
University of Denver
Denver, CO, USA
stephen.hutt@du.edu

Ryan S. Baker
University of Pennsylvania
Philadelphia PA USA
rybaker@upenn.edu

ABSTRACT

Massive Open Online Courses (MOOCs) have increased the accessibility of quality educational content to a broader audience across a global network. They provide access for students to material that would be difficult to obtain locally, and an abundance of data for educational researchers. Despite the international reach of MOOCs, however, the majority of MOOC research does not account for demographic differences relating to the learners' country of origin or cultural background, which have been shown to have implications on the robustness of predictive models and interventions. This paper presents an exploration into the role of nation-level metrics of culture, happiness, wealth, and size on the generalizability of completion prediction models across countries. The findings indicate that various dimensions of culture are predictive of cross-country model generalizability. Specifically, learners from indulgent, collectivist, uncertainty-accepting, or short-term oriented, countries produce more generalizable predictive models of learner completion.

CCS CONCEPTS

• Applied computing~Education~E-learning • Computing methodologies~Machine learning • Information systems~Information systems applications~Data mining

KEYWORDS

MOOCs, Generalizability, Cross-country, Cross-culture, Completion, Predictive Modeling, MORF

ACM Reference format:
Juan-Miguel Andres-Bray, Stephen Hutt and Ryan S. Baker 2023. Exploring Cross-Country Prediction Model Generalizability in MOOCs. In L@S '23: *Proceedings of the Tenth ACM Conference on Learning @ Scale* (2023), July 20–22, 2023, Copenhagen, Denmark. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3573051.3593380

1 INTRODUCTION

Over the last decade, Massive Open Online Courses (MOOCs) have opened e-learning materials from top institutions to a broader audience and increased the accessibility of quality educational content to a global network [1, 32]. They have allowed learners to learn at their own pace, in their own environment, across thousands of available courses. However, MOOCs have suffered from steep attrition rates since their inception [21]. In seeking to address this issue, researchers have studied how to support learner retention (e.g., [2, 29]), expressing a continued need for accurate prediction of learner outcomes and subsequent development of automated interventions.

Despite MOOCs having a worldwide audience, however, the majority of MOOC research has not accounted for the differences in learners' country of origin and cultural background. Studies have found that learners from Western, educated, industrialized, rich, and democratic (WEIRD) societies account for the majority of research subjects in psychology—96% based on a 2008 survey of the top psychology journals—while only accounting for 12% of the world's population [16]. Hence, researchers should consider how well their published findings generalize across country borders and cultures in order to support the needs of learners less represented in the literature.

In this paper, we are interested principally in generalizability across cultural groups at scale – does a model trained on one population perform just as well on another population? Further, what factors influence this generalizability? We are not the first to consider this area of research, indeed a recent study by Li and colleagues [27], for example, sought to investigate the generalizability of prediction models developed using survey data from the United States (a WEIRD country) to survey data gathered from learners from other countries. They found that models developed using US data could predict achievement in data from other developed countries with high accuracy, but that model performance dropped considerably for less developed countries. If this is also true for MOOC courses, then existing prediction models [2, 7, 10] developed predominantly with learners from a small number of countries may be less effective for learners from other countries. Several papers have raised questions about how broadly prediction models developed for MOOCs can generalize. However, most existing work has looked at generalization between course runs or different courses, rather than different national populations of learners [8, 26, 38]. Therefore, in this study, we explore the role of national cultural differences in the degree to which models of student success in MOOCs generalize, asking the research question *Are country-level cultural features*

L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07…$15.00
https://doi.org/10.1145/3573051.3593380

good predictors of whether prediction models generalize between countries?

We investigate these questions using models predicting course completion. In predicting this metric, it is important to recognize that not all learners enroll in MOOCs intending to complete. For example, some seek to gain just enough knowledge to publish in their field, join a community [39], or attain various job-related benefits [37]. Course completion, however, continues to be the most researched and widely used success metric for MOOCs. We conduct our experiments within a dataset of almost 2 million learners enrolled in the full 2012-2015 selection of Coursera MOOCs offered at the University of Pennsylvania. We leverage the MOOC Replication Framework [20] to conduct these analyses. We examine the impact of country-level cultural features on the generalizability of completion prediction models across diverse learner populations from 81 different countries. In addition, we identify which features and differences relate to the degree of generalizability seen. To our knowledge, this paper presents the broadest exploration yet into the role of country-level features on the generalizability and application of educational prediction models across countries.

1.1 Cross-Country Generalizability in e-Learning Research and MOOC Research

Investigations into the cross-country generalizability of findings have been rare across e-learning fields with some notable exceptions, such as a study by San Pedro and colleagues [31], which reported successful generalization of carelessness models between learners in the US and the Philippines. However, another study found that transferring models across learner populations led to poor model performance, relative to the training country's own baseline model performance. Specifically, help-seeking models for intelligent tutoring systems transferred to some degree between learners from the US and the Philippines, but not to Costa Rica [24], possibly due to the different ways students sought help outside the learning systems in these different countries.

MOOC scholarship has yet to investigate the issue of cross-country generalizability, a critical avenue of research given findings that country of origin is significantly related to how learners engage with MOOCs [14, 22, 28]. A study by Liu, Brown, and colleagues [28] found significant differences in learner interactions in a MOOC depending on the learner's country. The study identified learner profiles based on how they participated in the MOOC (e.g., those who predominantly only took quizzes, only watched videos, etc.), clustered the countries in their dataset based on several Hofstede's [17] cultural dimensions and found significantly different learner profile compositions per cultural cluster. Guo and Reinecke [14] found that the probability learners would interact with a MOOC in a non-linear manner (i.e., by navigating backward to a previous module instead of continuing on the sequence) varied based on their country of origin. Specifically, learners from countries with lower student-teacher ratios (e.g., the US and European countries) were significantly more likely to interact in a non-linear manner than those from higher student-teacher ratios (e.g., Kenya, India, etc.).

Ultimately, to better support all learners towards success, published findings in MOOC research need to generalize across different learner populations. This leads to this study's main research question: what country-level measures lead to better or worse generalizability in cross-country predictive modeling? As noted in a review by Baker and colleagues [5], despite a small number of examples (such as the ones given above), this question has not been systematically investigated by the field, and researchers still do not have a clear idea of what factors to look at. It may be possible to select factors for consideration based on studies that investigate the effectiveness of educational findings across different groups of students, such as socio-economic status [9], national wealth [24], or whether the student comes from a collectivist or individualist cultural background [23]. Identifying which measures relate or contribute to better (or worse) generalization of models across countries can help us ensure that the models we use for intervention are accurate and appropriate for the full variety of learners worldwide.

2 Data

2.1 MOOC Replication Framework (MORF)

This study used the MOOC Replication Framework (MORF [20]), a research platform developed to reduce technical, data, and methodological barriers to conducting replication studies on MOOCs. For reasons of security, privacy, and data ownership, the data available in MORF is not available for export or download, but instead is available for analysis through a secure platform governed by a data use agreement.

This study was conducted using learner data from MOOCs offered by the University of Pennsylvania. Only courses taught primarily in English were used in this study, as other courses tended to have learners from a smaller set of countries. In MOOCs during the time period studied, a course typically ran for a set number of weeks in which learners could enroll, engage in, and earn a completion certificate. Due to demand, some courses were offered multiple times. Each offering or instance of a course is referred to here as a session (as in [8, 13]). That is, each course could have been offered multiple times and thus have multiple course sessions, depending on how many times the course was offered over the period of time covered in the dataset. This dataset had a total of 45 courses; 27 of these courses were offered multiple times, resulting in a total of 98 course sessions. Full details of the courses included in the dataset can be seen in [4].

The volume of data within the framework allows for the investigation of research questions at scale, making it possible to determine what findings hold across different courses and iterations of those courses, and which findings are unique to specific kinds of courses and/or kinds of participants.

2.2 Measures of National Culture

To characterize each country in this analysis, we first consider measures of culture. This study considers two measures to operationalize national culture: Hofstede's cultural dimensions framework [18] and overall national happiness, as measured by the World Happiness Report [15]. The former is among the more commonly used cultural frameworks for investigating cultural differences in computer-based learning systems [5]. The latter, however, has never been used to investigate learning directly but has been used extensively to measure psychological well-being [25] and the conditions that support a person's continued drive to learn [15].

2.2.1 Hofstede's Cultural Dimensions. Hofstede's six cultural dimensions, outlined in Table 1, are used in this study to more closely examine cross-cultural variations within the learner sample. Dimension scores were developed from the survey responses of over 100,000 participants and are currently available online for 107 countries or regions[1].

Table 1. Overview of Hofstede's Cultural Dimensions [18]

Dimension	Description
Power Distance Index (PDI)	Measures the perception of power distribution. High-scoring cultures in this dimension denote a large power distance, where people tend to be deferential to figures of authority and accepting of unequal distributions of power. People from low power distance cultures readily question authority and expect to participate in decision making.
Individualism vs. Collectivism (IDV)	Within this measure, high-scoring cultures are considered individualistic characterized by a tendency to focus on their own needs and those of their immediate family. Low scoring cultures are considered to have a collectivist culture.
Gendered Role Index (GRI). Previously referred to as "Masculinity vs. Femininity"	Measures a cultures adherence to strict gender roles. A high score here implies a strictly gendered society.
Uncertainty Avoidance Index (UAI)	Measures the social tolerance for ambiguity and uncertainty. High-scoring cultures are uncertainty avoidant, and people in these cultures believe that uncertainty is a "continuous threat that must be fought."
Long-Term Orientation vs. Short-Term Normative Orientation (LTO)	Measures the inclination of a given culture to focus on future rewards. Cultures that score highly on this measure are considered long term oriented and value thrift and perseverance.
Indulgence vs. Restraint (IND) (added to the framework in 2010 [18])	Measures the social perceptions around human desires and gratification in comparison to regulation and strict social norms. A high score on this measure implies an indulgent culture that values leisure and personal control.

These dimensions have been widely cited across multiple disciplines, including psychology, sociology, education, and marketing [33–35]. They have been used to analyze and explain differences in various behaviors in educational technology. For example, in their study on help-seeking model transfer between countries, Ogan and colleagues [30] hypothesized that their mixed results and the apparent mediating role of student collaboration were due to differences among the three countries in Hofstede's cultural dimension on adherence to gender roles. Kizilcec and Cohen [23] investigated the efficacy of a self-regulation strategy between countries on opposite ends of Hofstede's individualism dimension and found that a strategy developed in Western countries significantly improved completion rates among learners from individualist countries (like the US, Australia, and France), but had no effect on learners from collectivist countries (like India, China, and Mexico).

2.2.2 Gross National Happiness. Another country-level metric considered is Gross National Happiness (GNH) or overall societal happiness, as reported in the World Happiness Report [15], an annual publication of the United Nations Sustainable Development Solutions Network. This report contains an index of national happiness based on a survey asking people to rate their satisfaction with aspects of their lives, such as their country's economy, social support, health, freedom to make life choices, generosity, and perception of corruption. The World Happiness Report publishes the estimated extent to which each of the six factors contributes to societal happiness. For this study, the GNH values used were from 2015 to match the final year of MOOC data used. This measure complements Hofstede's dimensions, bringing in not just culture but a key aspect of daily experience.,

2.3 Additional Country-Level Measures

In addition to Hofstede's cultural dimension indices and happiness index, we included four general country measures. These were: enrollment size (for the country) across all MOOCs in the data set (derived from data in MORF), National Population, Gross Domestic Product (GDP), and Per Capita GDP. The latter three were all taken from publicly available 2015 data[2] to be consistent with other measures.

[1] https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

[2] https://data.worldbank.org/indicator/

2.4 Sample Size

Our initial dataset comprised of over two million learners (*N*=2,008,618) from 118 countries. Learners from countries not present in either the Hofstede or Happiness databases of national variables were dropped from all analyses in the study (*N*=88,741; 4.42%), resulting in a dataset of over 1.9 million learners (*N*=1,919,877) across a total of 81 countries. Learners from the United States were the largest group of learners in the dataset (33%). To better contextualize this, the next most represented country, India, accounts for just 8% of the dataset.

3 Research Design

This study was divided into three phases. The first phase establishes the best-performing completion models per country. The second phase considers the *distance* between every country pair (i.e., a training country and a testing country) by comparing the cross-country model performance with the training country's own within-country, baseline model performance. Finally, the third phase seeks to explore the relationship between the cross-country distances and several country-level measures.

4 Phase 1: Within Country Models and Baseline Performances

The first phase of our experiment establishes a *within-country* baseline from which our generalizability analysis can be conducted. Put simply, we must examine how well a model trained on a single-country dataset performs on unseen members of that dataset, before we can evaluate how well it generalizes to another dataset. By modeling student outcomes within a country (i.e., all the learners from that country, across multiple MOOCs), we can examine how model performance varies by country, relative to cultural factors (described above) before considering a cross-country evaluation. For this purpose, we define a series of standard features that can be extracted for all course offerings and student identities to support both this, and future, phases.

4.1 Methodology

In order to assess cross-country generalizability, we first build predictive models of completion for each country. Doing so allows us to establish baseline model performances, i.e., model performance when trained and tested on a country's own data.

3.1.1 Data Cleaning and Feature Engineering. For each learner-session pairing, we first established if the learner completed that course (either regular completion or with distinction) and the learner's location while taking the course. The learners' IP addresses were used to geolocate their country, labeled using MaxMind's GeoIP2 Precision Country Service API3. In the cases where a learner used multiple IP addresses, the IP address that was used the most was the one attached to the learner. Dependencies, such as overseas territories, constituent countries, and Areas of Special Sovereignty or autonomous territories (e.g., Curaçao, constituency of the Netherlands; Puerto Rico, territory of the US) are labeled by GeoNames separately from their governing countries. As such, all analyses treated dependencies as separate from their governing countries.

To allow for a standardized analysis over time, sessions were divided into eight equal (within session) increments (relative to official start and end dates). Due to the varying length of sessions, increments ranged from 3.5 days (i.e., three days and 12 hours) to 11.375 days (i.e., 11 days and nine hours), with a median of 6.125 days and a standard deviation of 2.26. The start and end dates and times of these increments were used in conducting feature engineering.

In each course, learners used several resources, e.g., the discussion forums, quizzes, peer assessments, and lecture videos. The features listed in Table 2 were pulled per learner per increment, and then z-scored to account for the varying increment lengths.

Table 2: Incremental Features Used in Building Completion Prediction Models

Feature	Definition
Forum Views	Total number of clicks related to any forum activity (e.g., viewing, posting, commenting)
Quiz Views	Total number of clicks related to any quiz activity (e.g., viewing, answering, submitting)
Peer Assessment Views	Total number of clicks to any peer-assessment-related activity
Lecture Video Views	Total number of clicks related to any video lecture activity (e.g., playing, pausing, increasing video speed, etc.)
Days Active	Total number of days active
Forum Threads Started	Total number of forum threads started
Responses	Total number of responses to others' forum posts
Respondents	Total number of others' responses on one's own forum posts
Time Spent	Time spent (in seconds) in the forums, quizzes, peer assessment, and video lectures; actions with a computed duration of over one hour were treated as disengagement and excluded from the sum

3.1.2 Prediction Modeling. We trained three prediction models using the scikit-learn and xgboost libraries in Python: CART (Classification and Regression Trees), Random Forest, and XGBoost. Informal hyperparameter tuning was conducted for the RF and XGB classifiers in order to determine which value for n_estimators (how many trees will be used in training) was optimal for the problem. Hyperparameter tuning was conducted on data from three countries of different sizes: small (Mauritius, N=1,008), medium (Egypt, N=20,368), and large (United Kingdom, N=70,260) countries. Five values for n_estimators were tested per classifier: 100 (default), 300, 500, 700, and 900. The following

[3] https://www.maxmind.com/en/geoip2-precision-country-service

values were optimal across all three countries, feature sets, and increments: n_estimators=700 for Random Forest and n_estimators=100 for XGBoost. These values were applied throughout the rest of the study.

We trained models for each course increment, considering two feature sets: 1) increment-only: features from only the current increment (Nfeatures=13) and 2) appended: features from the current and all previous increments (Nfeatures=13 * increment number). Per combination, 10-fold cross-validation was conducted. Stratified sampling was used in assigning folds to preserve completion rates. A total of 480 models were trained and tested per country, ten (one per fold) for each combination of classifier (3), feature set (2), and increment (8).

We evaluated model performance using the Area Under the Receiver Operating Characteristic Curve (AUC ROC). An AUC ROC of 0.5 indicates chance level of performance, while a value of 1 means perfect classification. AUC ROC scores were averaged across each classifier-feature set-increment combination's respective ten folds. In order to determine each country's best performing model, averaged AUC ROC scores were compared across increments in each classifier-feature set combination using the statistical testing procedure from [12].

This was performed by iteratively comparing the AUC ROC of an increment with the AUC ROC of all future increments. This was conducted to determine if the model performance at an increment (e.g., at Increment 4, i.e., halfway through the course) was significantly lower than the model performance of future increments (e.g., increment 7, i.e., after about 87.5% of the course), which can be expected to have higher AUC ROC scores due to the higher amount of data available. If any comparison came out significant after conducting a Bonferroni correction [11], (e.g., the model performance at Increment 4 was significantly lower than the performance at Increment 7), then that increment was not used as the best performing model. Otherwise, if no comparisons came out significant (e.g., the model performance at increment 4 was not significantly lower than the performance at any of the future increments), then the model at that increment was treated as the best performing model. However, models requiring data from Increment 8 (i.e., the final increment) were dropped from consideration for two reasons : (1) Having to wait for data from the final increment of a course precludes stakeholders from conducting interventions, which is counterintuitive to the goal of predicting learner completion, and (2) Models that used the appended feature set in Increment 8 outperformed all other incremental models 100% of the time due to their use of the data of the entire course run.

The comparisons resulted in a final selection of six AUC ROC scores per country, one for each classifier and feature set combination. From here, the best performing completion prediction model was chosen per country, and its AUC ROC was treated in the subsequent analyses as the country's baseline model performance.

4.2 Results

We trained and evaluated a total of 81 models, one per country included in our analysis. For each model, all learner data from that country was used (i.e., across multiple MOOCs). Baseline AUC ROC scores (on test folds) across the 81 countries ranged from 0.874 (Iraq) to 0.992 (China), with a median of 0.979. The summary of a descriptive analysis reporting which model/feature combination was most successful for each country is shown in Table 3.

Table 3: Descriptive Results of the Parameters Used in the Best Performing Models

	Increment Only	Appended	Total
Random Forest	5	18	23 (28%)
XGBoost	5	53	58 (72%)
Total	10 (12%)	71 (88%)	81

Note. Parameters presented across the different combinations of classifiers (rows) and feature sets (columns). Each combination reports the number of countries whose best performing model used the respective combination and percentages of countries per parameter.

Out of the 81 best classifiers examined, 53 used the combination of XGBoost and the appended feature set (as seen in Table 3). Of those 53 classifiers, we examined which increment (e.g., data from how far through the course session) provided each result, shown in Table 4. As a reminder, increments span an eighth (i.e., 12.5%) of each course, where Increment 1 is the first eighth, Increment 2 is the second eighth, and so on. We observed that the majority of models used data until Increment 4 (i.e., until halfway through the course). However, countries with larger enrollment sizes benefitted from more data, as evidenced by the substantial leap in the mean enrollment size of countries needing data from either Increments 5 or 6. Still, the majority of the countries' models could predict learner completion using data until just Increment 4 (halfway through the course).

Table 4: Descriptive Results of the Increments used in the Best Performing XGB-appended Models

Increment	N Countries
1	1 (2%)
2	3 (6%)
3	4 (8%)
4	31 (58%)
5	6 (11%)
6	8 (15%)

Note. Each row reports the number of countries whose best performing XGB-appended model uses the respective increment, *N*=53.

3.2.1 Correlation Analysis. Nonparametric correlations were conducted between the countries' baseline model performances and the set of country-level measures. The Benjamini-Hochberg [4] post-hoc correction was used to control for the number of correlations conducted.

Enrollment size was significantly positively related with baseline model performance (rho=.880, p<.001); as enrollment size increased, so did baseline model performance. Country wealth was also strongly correlated with baseline model performance (GDP: rho=.765, p<.001; per capita GDP: rho=.480, p<.001). Happiness (rho=.354, p=.001) and cultural dimensions that look at individualism/collectivism (rho=.423, p<.001) and long-term/short-term orientation (*rho*=.480, *p*<.001) also had significant positive relationships with model performance; better-performing models were obtained for happier, more individualistic, and more long-term oriented countries. The full results can be found in Table 5.

Table 5: Correlation Results Between Baseline AUC ROC Scores and the Country-Level Measures

Measure	Correlation
Enrollment Size	0.880 *
Gross Domestic Product	0.765 *
Long-Term/Short-Term	0.480 *
Per capita GDP	0.466 *
Individualist/Collectivist	0.423 *
Happiness	0.354 *
Population	0.353 *
Uncertainty Avoidance	-0.093
Gendered Role Index	0.221
Power Distance	-0.219
Indulgence/Restraint	0.120

* p< .001 and significant after Benjamini-Hochberg [4] correction.

3.2.2 Regression Analysis. Linear regression was conducted to determine whether country-level measures were predictive of model performances. Two linear models were fit, the first using only the countries' six cultural dimension indices, and the second using only the remaining measures (i.e., happiness index, enrollment size, population size, GDP, and per capita GDP). Due to the high correlations between the country-level measures (Table 6), stepwise backward selection was conducted to account for collinearities and to remove suppression effects in both linear models using the step function in R's stats library. This function searches for the best possible regression model by iteratively selecting and dropping variables to arrive at a model with the lowest possible AIC (Akaike Information Criteria) [5].

Feature selection on the six cultural dimension indices revealed that long-term/short-term orientation (F(1, 78)=13.114, p<.01) and individualism/collectivism (F(1, 78)=4.806, p=.031) were most relevant to model performance. A model that regressed AUC ROC scores on indices from these two dimensions revealed that only long-term/short-term orientation significantly predicted model performance (β=.39, p<.001).

Feature selection on the country-level measures of happiness, wealth, and size revealed that happiness (F(1, 78)=20.123, p<.001) and population size (F(1, 78)=9.796, p=.002) were most relevant to model performance. A second model was regressed on these two measures and revealed that both were predictive of model performance within the full model (happiness: β=.51, p<.001; population: β=.31, p=.002).

5 Phase 2: Cross-Country Model Distances

Phase 2 considered how models trained in Phase 1 performed when classifying instances from countries other than the training country. For each possible pairing, we are then able to calculate the distance between AUC ROC scores, which can be used as a metric of how well a model has generalized. For example, we investigated how well a model trained on residents of the United States performs when evaluated on residents of the United Kingdom. These differences in model performance can then be used in conjunction with country-level measures to provide a deeper analysis of what impacts generalizability. For example, is a model trained on a country with a high GDP more or less likely to generalize than a model trained on a country with a low GDP?

5.1 Methodology

First, a list of all possible training and testing country pairs was compiled, resulting in a total of 6480 pairs (81 training countries * 80 testing countries). For each pair, one country was assigned to be train, the other test. Prediction modeling iterated over all train-test country pairs. In each iteration, the details of the training country's predictive model were pulled (i.e., feature set and increment) and applied to the testing country's dataset. Each of the training country's 10 fold-level models from (trained in Phase 1) were applied to the test country instances. This resulted in ten AUC ROC scores, which were averaged to determine the models' cross-country performance. Finally, *distances* between country pairs were computed by subtracting the cross-country AUC ROC score from the training country's baseline performance: $distance = AUC\ ROC_{baseline} - AUC\ ROC_{cross}$. Thus, a negative distance implies that that model performed better on the cross country, while a positive difference implying worse performance on the test country.

5.2 Results

Cross-country AUC ROC scores ranged from 0.747 (Iraq→Mauritius) to 0.993 (Brazil→Luxembourg), with a median of 0.973 across the 6480 country pairs. Distances ranged from -0.042 (Lebanon→Ethiopia) to 0.217 (Netherlands→Mauritius), with a median distance of 0.005 across the 6480 country pairs. Negative distances represent cases wherein the cross-country performance outperformed the training country's baseline model performance. For example, the performance of Lebanon's model on Ethiopia's data (*AUC ROC*=0.976) outperformed Lebanon's own baseline model performance (*AUC ROC*=0.935), resulting in a distance of -0.042. The distribution of distances can be found in Figure 1.

4.2.1 Correlation Mining. Nonparametric correlations were conducted between each training country's mean cross-country AUC ROC scores (raw scores not differences, e.g, how well did a

country's baseline model do on average when applied to the 80 other countries, regardless of the test country's baseline) and the training country's country-level measures (Table 6), using the Benjamini-Hochberg post-hoc correction for significance [6]. The training country's enrollment size (i.e., its number of training data points) was the most strongly correlated with mean cross-country model performance (*rho*=.846, *p*<.001), suggesting that, despite our hypothesis that differences in demographic and cultural factors lead to degraded model performance, models trained on countries with a large enrollment size are able to perform well on data from other countries. Measures of country wealth also strongly related to mean cross-country performance (GDP: *rho*=.732, *p*<.001; per capita: *rho*=.311, *p*=.005), suggesting that wealthier countries are also able to produce more generalizable models.

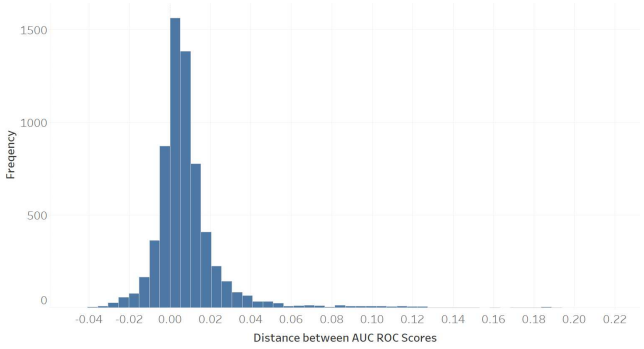

Figure 1: Distribution of Cross-Country Distances

Table 6: Correlation Results between Cross-country Model Performance and Training Country Country-Level Measures

Measure	Correlation
Enrollment Size	0.846 **
Power Distance	-0.019
Individualist/Collectivist	0.265 *
Gendered Role Index	0.320 **
Uncertainty Avoidance	0.035
Long-Term/Short-Term	0.304 **
Indulgence/Restraint	0.189
Gross Domestic Product	0.732 **
Happiness	0.201
Population	0.430 **
Per Capita	0.311 **

* *p*<.05, ** *p*<.001 and significant after Benjamini-Hochberg [4] correction.

6 PHASE 3: UNDERSTANDING MODEL DISTANCES

In this third phase of the study, we explore the relationship between the cross-country distances and the differences in the country-level measures in order to analyze how each measure relates to model generalizability. Whereas Phase 2 considered how country-level measures impacted model performance, Phase 3 considers how *differences* in country-level measures relate to the differences in model performance.

5.1.1 Correlation Mining. Nonparametric correlations were conducted between the cross-country AUC ROC distances and differences in the country-level measures (calculated using the same formula), using the Benjamini-Hochberg [6] post-hoc correction. Correlations were also conducted between distance and the absolute country-level measure differences in order to assess whether simply the presence of a difference mattered, or the direction of a difference mattered. The results of this analysis can be found in Table 7.

Table 7: Correlation Results between Cross-country Model Performance and Training Country Country-Level Measures

Difference In	Correlation with Difference	Correlation with Absolute Difference
Enrollment Size	0.016	0.050**
Power Distance	-0.249**	-0.010
Individualist/Collectivist	0.306**	-0.014
Gendered Role Index	-0.046**	-0.008
Uncertainty Avoidance	-0.057**	-0.019
Long-Term/Short-Term	0.288**	-0.006
Indulgence/Restraint	-0.128**	-0.009
Gross Domestic Product	0.033**	0.006
Happiness	0.208**	0.055**
Population	-0.142**	0.049**
Per Capita GDP	0.296**	0.009

** *p* < .001 and significant after Benjamini-Hochberg (1995) correction.

If the direction of a difference didn't matter—if just the presence of a difference mattered—then the absolute difference analysis would have resulted in a stronger correlation than the difference analysis. However, these results suggest that the direction of difference is more important than the absolute difference in these variables between countries (e.g., Figure 2), except for differences in enrollment size.

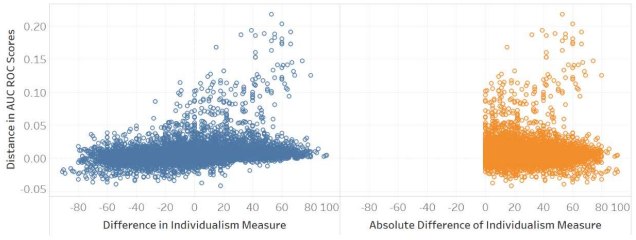

Figure 2: Graphical Representation of Distance x Difference and Absolute Difference in Measures of Individuality

Differences in power distance, adherence to gender roles, uncertainty avoidance, and indulgence were significantly

negatively correlated with cross-country model distances. These findings suggest that models trained on data from countries scoring high in these dimensions are likely to generalize (i.e., have a lower distance) on data from countries scoring low in the respective dimension, but not the other way around (e.g., indulgent country to restrictive country). Differences in happiness, individuality, and long-term orientation, on the other hand, were significantly positively correlated with model distance, suggesting that the *lower* in these dimensions the training country scored compared to a testing country, the more generalizable their models (e.g., less happy country to happier country).

5.1.2 Regression Analysis. In order to further investigate the relationship between the feature differences and the cross-country distances, regression analyses were conducted to measure the effects of each country-level measure difference on cross-country distance. Two linear mixed-effects models were fit on the country-pair dataset (*N*=6480) to estimate the effect of the cross-country measure differences on each pair's distance, with the pair's training country as the model's random factor. The results can be found in Table 8. The first model was regressed on differences related to Hofstede's six cultural indices. We performed backward elimination feature selection, which resulted in Gendered Role index being dropped from the model.

Table 8: Cross-Country Distance Regression Results

Predictors	*β*	*p*	*β*	*p*
(Intercept)	**-0.00**	<0.001	-0.00	<0.001
Power Distance	**-0.18**	<0.001		
Individualist/Collectivist	**0.08**	<0.001		
Uncertainty Avoidance	**0.07**	<0.001		
Long-Term/Short-Term	**0.18**	<0.001		
Indulgence/Restraint	**-0.07**	<0.001		
Enroll Size			**-0.14**	<0.001
Happiness			**0.08**	<0.001
GDP ($B)			**0.14**	<0.001
Population			**-0.03**	0.040
Per Capita			**0.18**	<0.001

In order to understand the relationships implied by the coefficients, Table 9 contains worked examples of four cases:

1. When the feature difference is positive and the coefficient is negative, the resulting effect on the predicted distance is a negative value, decreasing the distance, thus implying a more generalizable model from train to test country.
2. When the feature difference is negative and the coefficient is negative, the resulting effect on the predicted distance is a positive value, increasing the distance, thus implying a less generalizable model from train to test country.
3. When the feature difference is positive and the coefficient is positive, the resulting effect on the predicted distance is a positive value, increasing the distance, thus implying a less generalizable model from train to test country.
4. When the feature difference is negative and the coefficient is positive, the resulting effect on the predicted distance is a negative value, decreasing the distance, thus implying a more generalizable model from train to test country.

Differences in views on power distance and indulgence/restraint had significant negative effects on the cross-country distances, as in Table 9(1). This implies that as the training country ranked higher in either dimension (i.e., $index_{train}$ > $index_{test}$) and the country pairs' views of that dimension diverged (i.e., greater difference), the more generalizable the models were (i.e., the lesser the distance). In other words, these findings imply that models trained on learners from more indulgent countries or countries where a hierarchy of power is more accepted are likely to generalize on data gathered from their neighbors on the opposite end of the respective dimension (i.e., the more restrictive countries or countries that are more accepting of distributed power).

The opposite was true for the other three dimensions (e.g., Table 9(3)): as the training country ranked higher in either dimension and the country pair's views in that dimension diverged, the less generalizable the models were (i.e., the greater the distance). This finding implies that models generated using data gathered on learners from more collectivist, uncertainty-accepting, and short-term oriented are more likely to generalize to their respective counterparts, but not the other way around. Despite the statistical significance of the effects of these cultural index differences, however, they only explain a very small percentage of the variance in the cross-country distances, R^2=.101.

Table 9: Worked Examples for Negative and Positive Cross-Country Distance Regression Coefficients

	Train Val	Test Val	Diff Val	Coefficient	Effect on Predicted Distance
(1)	80	49	31	-0.18	-5.58
(2)	55	77	-22	-0.18	3.96
(3)	33	25	8	0.14	1.12
(4)	40	83	-43	0.14	-6.02

The second model was regressed on the other cross-country measure differences—differences in enrollment size, GDP, self-reported national happiness index, population, and per capita GDP. Despite high collinearity between features, all differences were included in the final model. Differences in enrollment size and population had significant negative effects on the cross-country distance. This implies that the more populous the training country was, or the more learners from the training country were enrolled compared to the test country, the more generalizable the models were. Conversely, the happier or wealthier the training country was compared to the test country, the less generalizable the models were. As in the Hofstede model, despite the statistical significance of the effects of these country-level measure

differences, they only explain a relatively small percentage of the variance in the cross-country distances, R^2=.067.

7 DISCUSSION

In this study, we examined how national-level variables impact the generalizability of predictive models in MOOC research. We did this by first determining baseline performance—training multiple models for each country and establishing which model performs best for each country. Next, we determined cross-country model generalizability by applying each country's best model on every other country in the dataset and comparing the results to baseline model performance (model performance on the training country). We then computed cross-country model distances as a metric of cross-country model generalizability using the baseline and cross-country AUC ROC scores. Distances were used to investigate the relationship between model generalizability and differences in various country-level metrics. These analyses found that models generally performed on par with their baseline model performances, only degrading by half a percentage point on average.

However, the degree to which models generalized across countries was significantly related to the differences in country-level measures of culture, happiness, wealth, and size. Hofstede's cultural dimensions were found to relate significantly to both the performance and generalizability of the completion prediction models. The study found that more individualistic or more long-term oriented countries were more likely to have better-performing within-country (baseline) models. It is worth noting that both these cultural indices were significantly positively correlated to the country's GDP and enrollment size, suggesting that individualistic or long-term oriented countries were also likely to be wealthier and have a larger MOOC presence (i.e., larger training dataset).

Further, differences in cultural views relating to power distribution, indulgence, individualism, and long-term orientation were significantly related to model generalizability. In the case of the indulgence dimension, for example, models trained on a more indulgent country (like Mexico or Sweden) generalize better on a more restrictive country, but caution should be placed when generalizing models trained on a more restrictive country. Ultimately, the findings suggest that training models on more power distant (e.g., China, the Philippines), more indulgent, more collectivist (e.g., Guatemala, Panama), or more short-term oriented countries (e.g., Ghana, Nigeria) was more likely to produce generalizable models. Countries that fit this profile, scoring high across all four dimensions, include Venezuela, Mexico, Ghana, and Nigeria, all of which have mid-range enrollment (mean=12627) and GDP (mean=$5.1B). On the other end are countries like Estonia, Lithuania, Latvia, and Hungary, which have both low enrollment (mean=4262) and GDP (mean=$0.5B). Both groups of countries have similar average baseline model performances, *AUC ROC*=0.97.

Gross National Happiness, or self-reported nation-level happiness, as measured by the World Happiness Report [11], was also related significantly to model performance and generalizability. Interestingly, while happiness was found to positively impact model performance (both within and cross-country), the difference in happiness between countries had an inverse relationship with model generalizability. That is, the happier a training country is compared to a testing country, the less generalizable the models. The relationship suggests that models produced using data from low-happiness countries were more likely to generalize compared to models produced using data from happier countries.

Finally, measures of wealth and size were also found to relate significantly to both model performance and generalizability. GDP, per capita, population, and enrollment size were all significantly related to within-country model performance, suggesting that larger, wealthier countries with a larger MOOC presence were likely to produce better-performing models. This finding is intuitive—larger and wealthier countries are likely to have more learners enrolled in MOOCs (as evidenced by significant correlations between these measures), and a standard principle in machine learning states that having a larger training data set ensures better model performance. Likewise, differences in these features all had significant effects on model generalizability. The relationship between generalizability and differences in size metrics—population and enrollment size—suggests that the larger the training country is compared to the testing country, the more generalizable the training country's model is. The findings related to differences in wealth, on the other hand, suggest that the wealthier the training country is compared to a testing country, in either GDP or per capita GDP, the less likely its model will generalize (i.e., higher positive difference in GDP or per capita suggests higher distance score).

However, despite the statistical significance of the effects of these country-level measure differences, they only explain a relatively small percentage of the variance in the cross-country distances. A likely explanation is that a number of other country-level factors are at play, ones not considered in this study. Perhaps Hofstede's (2010) cultural dimension framework is insufficient in fully describing cultural differences across, or even within, countries (as explained in the Limitations section below). Perhaps other access or socioeconomic differences not accounted for in this study are also contributing to the model distances. What is clear, however, from these results that cultural differences do impact learning in MOOCs. We can use the results presented here as a step towards more culturally sensitive pedagogy for online learning, revising the "one size fits many" model used in a number of MOOCs.

7.1 Limitations and Future Work

The findings from this study serve as an initial examination of the relationships and patterns across countries as they relate to MOOC analytics. Our methods and results serve as a starting point for additional analyses and further refinement for future

cross-country analyses. For example, our initial feature set contained low-level features that are commonly used in MOOC research [3, 8, 13], however, there is potential for additional feature engineering and processing. Further features should be considered, to fully understand the factors that are predictive of model generalization. Similarly, future work can consider how the time increments impact results and potential variations by country and cultural variables. Future work should also conduct deeper investigations into why and how these country-level measures affect model generalizability, perhaps using interview methods to probe these relationships further.

As noted above, the study was limited by the type of success metric investigated in the training and testing of predictive models. MOOC scholarship has evolved from investigating course completion as the sole metric of learner success—learners have been found to come into these courses with varied goals and motivations, e.g., publishing or joining a professional organization in the same field [39], or attaining various job-related benefits [37]. Future work should replicate this approach across additional success metrics, and examine if the cultural moderators vary depending on the success metric considered.

Our study was also limited by the metrics used to quantify culture. A review by Baker and colleagues [5] differentiates between macro- and micro-theories of culture. Macro-theories of culture attempt to "categorize all groups in the world according to some number of cultural dimensions" (p. 2). Hofstede's cultural dimension framework falls into this category of cultural theories, in addition to other widely-cited frameworks: the Model of National Cultural Differences [36] and the nine dimensions presented in the GLOBE study [19] . Micro-theories on culture, on the other hand, seek to contextualize culture down to the individual-level. In these theories, culture is "embedded in particular actors' specific practices and activities that take place in particular contexts" [5]. They place an emphasis on a subject's own cultural identity. However, because micro-theoretical approaches to culture are limited in their generalizability [5] , and because this granularity of data would again be difficult to gather at the scale MORF operates on, our study was limited to macro-views of culture—specifically Hofstede's cultural dimensions.

8 Conclusion

The methods used in this study provide a novel approach to examining cross-country prediction model generalization. Understanding what, why, and how factors lead to generalization of predictive models between countries will not only lead to better informed culturally-sensitive pedagogy for learners around the world, it will also lead to a new and deeper understanding of how culture influences learner-computer interaction. In the meantime, the implications from the findings of this paper are clear: researchers developing and studying predictive models in MOOCs need to start accounting for differences in learner country.

ACKNOWLEDGMENTS

We would like to thank the MOOC Replication Framework for allowing us to use their infrastructure and data for this project, and Chelsea Porter for assistance in paper formatting and preparation.

This research was supported by the National Science Foundation (NSF) (NSF-OAC#1931419). Any opinions, findings, and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Adamopoulos, P. 2013. What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses. *Thirty Fourth International Conference on Information Systems*. 2013, (2013), 1–21. DOI:https://doi.org/10.1145/1164394.1164397.

[2] Alamri, A. et al. 2021. MOOC next week dropout prediction: weekly assessing time and learning patterns. *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17* (2021), 119–130.

[3] Andres, J.M.L. et al. 2017. Replicating 21 findings on student success in online learning. *Technology, Instruction, Cognition, and Learning*. 10, 4 (2017), 313–333.

[4] Baker, R. et al. 2022. Research Using the MOOC Replication Framework and E-TRIALS. *2022 IEEE Learning with MOOCS (LWMOOCS)* (2022), 131–136.

[5] Baker, R.S. et al. Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning In Context*. 1, 1, 1–13.

[6] Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. 57, 1 (1995), 289–300. DOI:https://doi.org/10.2307/2346101.

[7] Brinton, C.G. and Chiang, M. 2015. MOOC performance prediction via clickstream data and social learning networks. *Proceedings - IEEE INFOCOM*. 26, (2015), 2299–2307. DOI:https://doi.org/10.1109/INFOCOM.2015.7218617.

[8] Brooks, C. et al. 2015. Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs). *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (2015), 245–248.

[9] Buolamwini, J. and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency* (2018), 77–91.

[10] Crossley, S. et al. 2017. Predicting success in massive open online courses (MOOCs) using cohesion network analysis. *Proceedings of the International Conference on Computer-Supported Collaborative Learning* (2017), 103–110.

[11] Dunn, O.J. 1961. Multiple comparisons among means. *Journal of the American statistical association*. 56, 293 (1961), 52–64.

[12] Fogarty, J. et al. 2005. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *Proceedings of Graphics Interface 2005* (2005), 129–136.

[13] Gardner, J. et al. 2018. Replicating MOOC predictive models at scale. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (2018), 1–10.

[14] Guo, P.J. and Reinecke, K. 2014. Demographic differences in how students navigate through MOOCs. *Proceedings of the first ACM conference on Learning@ scale conference* (2014), 21–30.

[15] Helliwell, J.F. et al. eds. 2015. *World Happiness Report 2015*. Development Solutions Network.

[16] Henrich, J. et al. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*. (2010). DOI:https://doi.org/10.1017/S0140525X0999152X.

[17] Hofstede, G. 1986. Cultural differences in teaching and learning. *International Journal of intercultural relations*. 10, 3 (1986), 301–320.

[18] Hofstede, G. et al. 2010. *Culture and organizations: software of the mind, intercultural cooperation and its importance for survival*. McGraw Hill.

[19] House, R.J. et al. 2004. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

[20] Hutt, S. et al. 2022. Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology*. 53, 4 (2022), 756–775. DOI:https://doi.org/10.1111/bjet.13231.

[21] Jordan, K. 2014. Initial trends in enrolment and completion of massive open online courses. *International Review of Research in Open and Distributed Learning*. 15, 1 (2014), 133–160.

[22] Kizilcec, R.F. et al. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *Proceedings of the third international conference on learning analytics and knowledge* (2013), 170–179.

[23] Kizilcec, R.F. and Cohen, G.L. 2017. Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences*. 114, 17 (2017), 4348–4353.

[24] Kulik, J.A. and Fletcher, J.D. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*. (2016). DOI:https://doi.org/10.3102/0034654315581420.

[25] Lan, X. et al. 2019. Parental autonomy support and psychological well-being in Tibetan and Han emerging adults: A serial multiple mediation model. *Frontiers in Psychology*. 10, (2019), 621.

[26] Le, C.V. et al. 2018. Communication at scale in a MOOC using predictive engagement analytics. *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19* (2018), 239–252.

[27] Li, X. et al. 2021. On the limits of algorithmic prediction across the globe. *arXiv preprint arXiv:2103.15212*. (2021).

[28] Liu, Z. et al. 2016. MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis. *International Educational Data Mining Society* (2016).

[29] Moore, R.L. and Wang, C. 2021. Influence of learner motivational dispositions on MOOC completion. *Journal of Computing in Higher Education*. 33, 1 (2021), 121–134.

[30] Ogan, A. et al. 2015. Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*. 25, (2015), 229–248.

[31] San Pedro, M.O.C.Z. et al. 2011. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. *Artificial Intelligence in Education* (2011), 304–311.

[32] Shah, Dhawal 2019. Online Degrees Slowdown: A Review of MOOC Stats and Trends in 2019. *Class Central*.

[33] Sndergaard, M. 1994. Hofstedes Consequences: A Study of Reviews. *Citations and*. (1994).

[34] Soares, A.M. et al. 2007. Hofstede's dimensions of culture in international marketing studies. *Journal of business research*. 60, 3 (2007), 277–284.

[35] Steenkamp, J.-B.E. 2001. The role of national culture in international marketing research. *International marketing review*. (2001).

[36] Trompenaars, F. and Hampden-Turner, C. 2011. *Riding the waves of culture: Understanding diversity in global business*. Nicholas Brealey International.

[37] Trumbore, A. 2021. Learner behavior and career benefits in business massive open online courses. *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021*. (2021).

[38] Veeramachaneni, K. et al. 2013. Moocdb: Developing data standards for mooc data science. *AIED 2013 workshops proceedings volume* (2013).

[39] Wang, Y. and Baker, R. 2018. Grit and intention: Why do learners complete MOOCs? *The International Review of Research in Open and Distributed Learning*. 19, 3 (2018)