**International Journal of AI in Education manuscript No.**
(will be inserted by the editor)

# Supporting College Choice Among International Students through Collaborative Filtering

**Caitlin Tenison · Guangming Ling · Laura McCulla**

Received: December 8, 2021 / Accepted: August 2, 2022

**Abstract** In this paper we use historic score-reporting records and test-taker metadata to inform data-driven recommendations that support international students in their choice of undergraduate institutions for study in the United States. We investigate the use of Structural Topic Modeling (STM) as a context-aware, probabilistic recommendation method that uses test-takers' selections and metadata to model the latent space of college preferences. We present the model results from two perspectives: 1) to understand the impact of TOEFL score and test year on test-takers' preferences and choices and 2) to recommend to the test-taker additional undergraduate institutions for application consideration. We find that TOEFL scores can explain variance in the probability that test-takers belong to certain preference-groups and, by accounting for this, our system adjusts recommendations based on student score. We also find that the inclusion of year, while not significantly altering recommendations, does enable us to capture minor changes in the relative popularity of similar institutions. The performance of this model demonstrates the utility of this approach for providing students with personalized college recommendations and offers a useful baseline approach that can be extended with additional data sources.

**Keywords** Recommender Systems · Collaborative Filtering · International Education · Undergraduate Education · Structural Topic Modeling

C. Tenison
Educational Testing Service
Tel.: +1 609-252-8348
E-mail: ctenison@ets.org
DOI: https://doi.org/10.1007/s40593-022-00307-0
*Preprint submitted to International Journal of Artificial Intelligence in Education*

## 1 Introduction

The beneficial impact of international higher education extends to the global, the national, and the individual level. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) targets and tracks international education, citing that international student exchange is a key mechanism for promoting long-term goodwill between nations (UNESCO, 2018). Within the United States (U.S.), the flow of international students makes higher education one of the country's largest exports in the service sector. In 2018, international students contributed more than $45 billion to the U.S. economy (Institute of International Education, 2021). Studying internationally improves student's job opportunities after graduation within the U.S. (Arbeit & Warren, 2013) and is linked to increased wages in their home countries (Arbeit & Warren, 2013; Guo, Zhang, & Ye, 2019; Wiers-Jenssen & Try, 2005). Despite these benefits, the rate of international student enrollment in the U.S. has been declining since the 2014-2015 academic year and recently has been especially hard hit by the Coronavirus pandemic (Institute of International Education, 2020). While some of the factors driving this decline are difficult to control, universities are looking for new ways to reach and recruit international students that will be good matches for their institutions.

Student-institution matching is a complex, multi-stakeholder problem. Currently there is considerable interest in university application decision-making and the criteria admissions officers use when selecting students (Hossler et al., 2019). This decision has important implications to issues of diversity and access to higher education and higher future income(Posselt & Grodsky, 2017). As a result, universities are exploring different types of admissions procedures to take into account the applicant's opportunities and experiences alongside measures of academic achievement (Bastedo, 2021). Before a university can make the decision of who to admit, students must first select that university and submit their application. Even if a student is a great fit for an institution, if they do not first identify that institution as an option this opportunity will remain closed (Black, Cortes, & Lincove, 2015). This puts pressure on students to make well-informed and well-researched choices. This is especially difficult for international students who are limited in their awareness of the institutions within the U.S.

In the current paper we explore a 'student-centered' approach to support institution-to-student matching that involves modeling students' preferences and identifying institutions that best align with those preferences. We frame the problem in terms of matching students to institutions that would appeal to them. This is a common framing within Recommender Systems research in which the focus is on the connection between users and relevant items. While there are many different approaches to recommendation, we propose a novel application of Structural Topic Modeling (STM) as a hybrid collaborative filtering approach to recommendation. This content-aware probabilistic recommendation method allows us to model the impact of student-level factors on choices while also treating student preferences as latent variables. We apply

this approach to model student preference using information contained within the registration and score reporting logs of the globally administered Test of English as a Foreign Language (TOEFL) internet based test (iBT). Using this data we consider several models of student preference that are sensitive to the previous research on international student application behavior (Section 2.1). We also explore how this type of model performs as a basis for recommending institutions to students.

Our research makes the following contributions to the field. First, this study extends the limited body of research on the application of Recommender Systems for supporting institution-to-student matching. While the scalability of prior approaches has been limited by the need for detailed information about both students and institutions, our results suggest that a hybrid recommendation approach that leverages student choice data can provide recommendations that are sensitive to preferences for specific areas of study, geographic locations and types of institutions. Second, this study demonstrates how STM can be used as a hybrid recommendation approach that allows the integration of covariates into the estimation of latent preference groups. This approach to recommendation supports the introspection into how certain features influence the expression of these latent groups. We demonstrate how this approach can be used to model and understand the factors that influence student preferences. This descriptiveness has valuable implications for the development and the transparent use of such algorithms within recommender systems (Tintarev & Masthoff, 2015). Third, our research provides insight into the role of different covariates on international student's preferences for schools within the U.S. We find that students preference is sensitive to their TOEFL scores suggesting that students are adjusting their preferences based on this information. On the other hand, our investigation year as a covariate indicates that the preference groups we identified are relatively stable in the face of the change in the U.S. government administration. Finally, much of the research on student-institution matching focuses on the admissions decision from the perspective of the institution. Identifying methods to support students in identifying schools has important implications for creating a diverse pool of qualified applicants for each institution. In presenting this case study we highlight one avenue by which we can support student's application decision making by modeling their preferences across the space of institutions.

The rest of the paper is organized as follows. First, we survey relevant research on the decision making and preferences of international students, and efforts to build recommender systems to support their choices. We then introduce STM as an approach for modeling student preference. Finally, we present a case study in which we apply STM to international student's score reporting behavior to capture how student's TOEFL score and year influence student's predicted preferences across U.S. institutions.

## 2 Background and Motivation

### 2.1 International Student Decision-Making and Preference

Considerable research has been conducted to understand the factors that influence a students' decision to study internationally and shape their preferences for different universities. These factors contribute to multiple decisions such as the country, state, and institution to study in (Nicholls, 2018) and the influence of these factors varies depending on the stage of a student's application from initial exploration of the space of options, to narrowing on a set of institutions to apply to, to finally choosing among admission offers (Hemsley-Brown & Oplatka, 2015). In this paper we focus on the initial exploration of the universities within the U.S. and student's choice of which institution to apply to. This decision has large downstream implications for the type of choices students have for enrollment and the experience they have when studying abroad. The research on international [1] students' application decision making processes relies on surveys, case-studies, and interviews of students who are either from the same country or have applied to the same institution (Abunawas, 2014; Mazzarol & Soutar, 2002; Nicholls, 2018; Zhu & Reeves, 2019). Models of the student decision-making process is captured as an interaction between multiple student-level and institution-level factors (e.g. Alfattal, 2016; Mazzarol & Soutar, 2002). One such model, the Push-Pull model, formalizes insights from years of research on international application decision-making (Mazzarol & Soutar, 2002). This theoretical model distinguishes between factors that 'push' students to seek out international institutions and factors that institutions use to attract, or 'pull', students. 'Push' factors in this model include factors such as a student's educational goals, financial constraints, and personal recommendations. 'Pull' factors in this model include factors such as reputation of the institution, degree program, cost issues, proximity to social support networks, and environment. This model additionally considers the impact of general factors, such as home country, and hypothesizes differences in undergraduate versus graduate students' decisions.

As with domestic students, institution reputation plays a large role in students considering universities (Nicholls, 2018). If a student does not already know about a university, finding easily accessible information geared towards the unique situations of international students can be challenging. As a result, larger universities engage in broad marketing campaigns to influence perceptions abroad (Chen, 2008; Wu & Naidoo, 2016). For smaller universities or universities looking to target a specific student population, the use of local education 'agents' to promote, recruit, and match students and universities is an increasing trend (Heuser, Martindale, & Lazo, 2016). While these agents can provide a personalized experience for students, such as coaching their application and guiding their choices, there are numerous incentive structures within this matching process that encourage fraudulent and exploitative be-

[1] The research we cite in Section 2.1, was conducted in the United States, Australia, and the United Kingdom, and views international students as students from other countries.

haviors that do not put students' best interests first(Flaitz et al., 2003; Hallak & Poisson, 2007; Heuser et al., 2016).

The prior research on the role of preference in international student application decision making has primarily focused on building theoretical models to identify preferences and understanding the role of different factors influencing those preferences. To our knowledge, little effort has been made to build formal or statistical models of this process. This comes as little surprise, since modeling these factors would require student data from multiple countries and require the tracking of applications across multiple institutions. Rather, most studies have focused on student choice within a single university (Nicholls, 2018) or limited students to a specific country or region (Zhu & Reeves, 2019). The information that Educational Testing Service (ETS) collects in administering the TOEFL exam affords a unique opportunity to model international students' applications decision-making process across a wide range of institutions while also accounting for factors such as students' English language skills as indicated by TOEFL score and the year the students sent their scores to U.S. institutions. We propose a student-centered approach to support matching by focusing on modeling students' preferences for universities and offering relevant options to consider.

## 2.2 Recommending Institutions to Students

In an effort to model international student application decision-making, we frame the problem in terms of matching students to institutions that would appeal to them. This is a common framing within recommendation research in which the focus is on the connection between individuals and relevant items (such as music, movies, restaurants). The type of approach used to build a recommender system depends on the availability of data. Content-based approaches cluster users and/or items based on their attributes, assuming that users with similar attributes will share similar preferences and items with similar attributes will be equally preferred. This approach benefits from a rich set of features about users and items. This can be problematic in cases where feature information is limited or incomplete. An alternative approach, Collaborative Filtering (CF), makes predictions about users' preferences based on their ratings of a few items and the past ratings of all users. Traditional CF approaches leverage users' ratings of items along some type of scale. Adaptations of the CF approach have extended to binary, positive-only data (Verstrepen, Bhaduriy, Cule, & Goethals, 2017). These adaptations support the application of this approach to explicit feedback such as 'likes', or to implicit feedback such as what items are selected (e.g., what institutions students apply to). The benefit of CF is that it requires no content information about the user or item, only the choices individuals make. Hybrid Approaches describe general class of approaches that takes advantage of both content-based features and collaborative filtering rating information to identify preferences. There are a variety

of hybrid models in the field, constructed to take advantage of different types of data and tailored to suit the needs of the specific domain of use.

Previous efforts to build recommendation systems for matching students to educational programs have focused on generating student profiles from application materials and self-reported preferences (Bokde, Girase, & Mukhopadhyay, 2015; Iyengar, Sarkar, & Singh, 2017; Ragab, Mashat, & Khedra, 2012). With the focus on features, these approaches favor content-based and neighborhood based collaborative filtering. Both Iyengar et al. (2017) and Ragab et al. (2012) provide students with recommendations based on the similarity between their application profiles and those of students already attending those universities. This limits how well these models can account for student specific preferences and constraints. Ragab et al. (2012) addressed this gap by authoring extensive decision rules to further govern the recommendation process. Bokde et al. (2015) proposed a means of learning these preferences using multi-criteria collaborative filtering that, although data driven, require a large sample of students to rate multiple universities across multiple criteria.

All three of these approaches share a reliance on extensive data about the students and a narrow focus on a small set of institutions. Extending any of these approaches to support recommendations for international students applying to U.S. institutions would face several challenges. First, the diversity of programs and student backgrounds reduces the set of common features available across all students, making content-based methods less informative. Additionally, we neither know the final university choice the students made nor how they performed once there, precluding the use of the approaches used by Iyengar et al. (2017), Ragab et al. (2012) and colleagues. Finally, the prior research has focused on building systems to support a narrow set of students and institutions. With over 2,000 higher learning institutions in the U.S. and the average student only applying to 7.4 (SD 5.2) institutions, the data we are modeling are much sparser than what prior approaches have used. Ultimately, we need a new approach that indirectly models the preferences of students from noisy, sparse data, but also provides a means to introduce features about the student.

## 3 Structural Topic Modeling for Recommendation

The current study explores how hybrid collaborative filtering can be used to generate recommendations that could help international students identify U.S. institutions to apply to based on their interests. We present the use of Structural Topic Models (STM) as a method for modeling student preferences from existing TOEFL score reporting data and metadata about the student. Traditional CF models infer user preferences from positive interactions between users and items. In the case of this work, we have access to information about what institutions students reported their TOEFL scores to. We regard score reporting as an indication that students intend to apply to that institution, this is treated as an indication of preference. Therefore, we treat student's

score reporting behavior as a type of binary, positive-only data. We represent this data as a matrix where each row is a unique student and each column is a unique institution. The matrix is populated with binary values indicating which institutions each student reported their scores to. Applying a CF approach, we use matrix factorization to compute, for every student-institution pair, the probability that students will prefer that institution.

### 3.1 Latent Dirichlet Allocation to Support Collaborative Filtering

In their matrix factorization framework, Verstrepen et al. (2017) describe a wide variety of factorization models that researchers have used for binary, positive-only data, such as ours. One basic factorization model used in CF is Latent Dirichlet Allocation (LDA). While this method is commonly used in topic modeling, applying LDA to recommendation makes the following assumptions:

1. Users belong to multiple preference-groups. Members of these groups select similar items, but the groups themselves are not directly observed.
2. User's choices are probabilistic events and they will select items based on group-item probability distributions.

In their paper describing an LDA-based recommendation system, Xie, Dong, and Gao (2014) describe these preference groups in terms of the many identities an individual may have. For example, we may have a student who is a young woman, from China, an aspiring artist, with relatives that live in California. Each of these identities have different preferences and will result in different choices (e.g. schools in California, schools with specialized Art programs, women's colleges). The student's final preferences are the probabilistic combination of these identities.

As a factorization model in a recommender system, LDA is used to calculate the probability $p(w|d)$ that a student, $d$, will prefer a specific institution, $w$, as a mixture of $z$ probability distributions induced by the hidden preference-groups. These hidden preference-groups are described as 'Topics', in the original terminology describing LDA (Blei, Ng, & Jordan, 2003). In LDA, the probability that a student belongs to a preference-group can be expressed as $p(z|d)$ and the probability that a preference-group will select institution $w$ is $p(w|z)$. A basic LDA model (e.g. Verstrepen et al., 2017; Xie et al., 2014) would represent this as follows:

$$p(w|d) = \sum_{z} p(z|d)p(w|z) \tag{1}$$

There are many methods for fitting LDA models to data including, Markov chain Monte Carlo, gradient descent and variational inference (Verstrepen et al., 2017). Many of these approaches operate by maximizing the likelihood of the model given the data. Additionally, there are many extensions of LDA, such as Correlated Topic Models (CTM) which use a logistic normal distribution instead of a Dirichlet to model topic proportions (Blei & Lafferty, 2007).

### 3.2 Incorporating Covariate Information into Collaborative Filtering Models

The LDA factorization model does not directly account for any metadata about the student. This model infers identities from the choices students make rather than using any student-level features. STM extends the basic LDA and CTM models by estimating $p(z|d)$ and $p(w|z)$ as functions of student-level covariates rather than as global parameters (Roberts, Stewart, & Airoldi, 2016; Roberts, Stewart, & Tingley, 2019). STM allows the introduction of two additional assumptions to traditional LDA:

1. Metadata about a user impacts the probability that they belong to a specific preference-group.
2. Metadata about the user can explain how preference-groups are associated with the probability of specific choices.

Roberts et al. (2019) refer to these assumptions as topical prevalence covariates and topical content covariates respectively. *Topical prevalence covariates* allow us to capture the predicted effect of student-level information on their membership to preference-groups. For example, two students applying to Ivy League institutions may seem similar; however, accounting for their TOEFL scores may distinguish a competitive applicant from someone applying to these institutions as a 'moonshot'. *Topical content covariates* capture how metadata about the student applying to different institutions can impact the probability that preference-groups are associated with the specific institutions. For example, institutions differ in their international outreach efforts which can result in regional differences in students' familiarity with institutions (Heuser et al., 2016). This might result in students from different countries belonging to the same preference-group but applying to different institutions.

The flexibility of STM to account for these covariates as part of the identification of latent groups makes it a popular tool for social science research (Roberts et al., 2016); however, we are not aware of any work applying STM as the factorization model within a CF approach to recommendation. In it's original application STM is used as a generative model that assumes that the words in each document arises from a mixture of *topics*. In our application we considering the institutions chosen by each student as arising from a mixture of the *preference-groups* students belong to. For the remainder of this paper, we refer to the latent preference-groups we identify with STM as 'topics', using the terminology specific to these types of models.

We provide a graphical illustration of STM using plate notation in Figure 1. In the following section we describe the generative process of modeling each student, $d$, with an STM model with $K$ topics and both topical prevalence and content covarates.

**Topical Prevalence Covariates**. STM captures $\theta_d$ as conditionally dependent on the topics $\Sigma$, the topical-prevalence covariates $X$, and their coefficients $\gamma$. The student-topic distribution, $\theta_d$, is estimated from a logistic-normal generalized linear model based on a vector of student covariates $X_d$ for each student $d$.

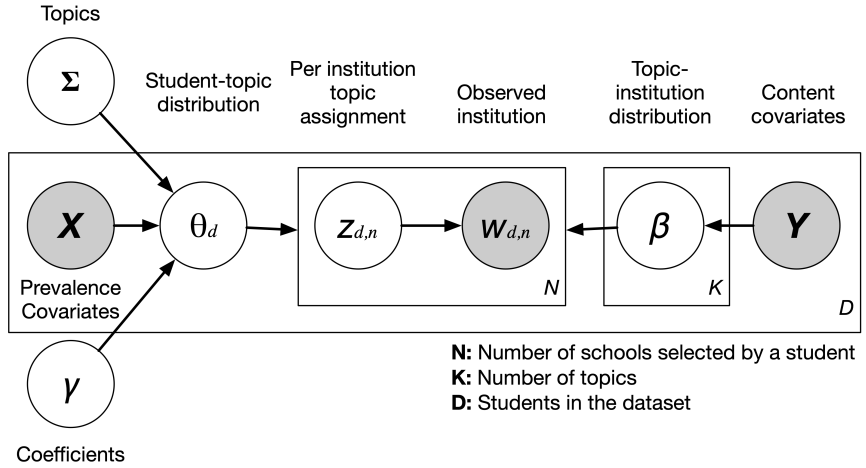


**Fig. 1** Graphical illustration of the structural topic model. Grey nodes indicate observable variables. We use the term 'topic' to refer to the latent-preference groups estimated by the model.

$$\overrightarrow{\theta_{\mathrm{d}}}|X_d\gamma, \Sigma \sim LogisticNormal(X_d\gamma, \Sigma) \tag{2}$$

Where $\gamma$ is a matrix of coefficients relating the student covariate values to the topic prevalence and $\Sigma$ is the topic variance-covariance matrix. In this formulation, $X_d\gamma$ is the matrix product of $X_d$ and $\gamma$.

**Topical Content Covariates**. STM captures the relationship between institutions and topics, in terms of the topic-institution distribution, $\beta_{d,k}$.

$$\beta_{d,k} \propto exp(m + k_k^{(t)} + k_{y_d,k}^{(c)} + k_{y_d,k}^{(i)}) \tag{3}$$

Such that for a given student-level content covariate $y_d$, we form the student specific distribution over institutions representing each topic $k$ using the baseline institution distribution $m$, the topic specific deviation $k_k^{(t)}$, the covariate group deviation $k_{y_d,k}^{(c)}$ and the interaction between the two $k_{y_d,k}^{(i)}$. Where all four terms are vectors containing one entry per institution in the set of institutions the student selected. Note, the variables $c$ and $i$ represent the specific content covariates and interactions, respectively.

**Combining Covariates**. For each institution chosen by a student, ($n \in 1, ....N_d$) we can draw that institution's topic assignment based on the student-specific distribution over topics. Given $\theta_d$, for each institution that student $d$ chooses, a topic, $z_{d,n}$ is sampled from a multinomial distribution:

$$z_{d,n}|\overrightarrow{\theta_{\mathrm{d}}} \sim Multinomial(\overrightarrow{\theta_{\mathrm{d}}}). \tag{4}$$

Conditional on the topic chosen $z_{d,n}$ and the topic-institution distribution $\beta_{d,k}$, we can draw an observed institution $w_{d,n}$ from that topic as follows,

$$w_{d,n}|z_{d,n}, \beta_{d,k=z_{d,n}} \sim Multinomial(\beta_{d,k=z_{d,n}}). \quad (5)$$

Using the STM package (Roberts et al., 2019), we use variational expectation-maximization to estimate the model parameters. Further technical details about the estimation of these parameters can be found in Roberts et al. (2016).

## 4 Case Study

As a case study of the use of STM to support institution-to-student matching, we model the latent preference-groups present within TOEFL score reporting data. We focus on two sources of information to capture as covariates in our model: the TOEFL score of test-takers and the year in which the test was reported. Besides providing us with a means to explore the influence of these covariates on test-takers' preferences, we consider how the incorporation of these factors alters the recommendations our recommendation system provides. We perform several offline evaluations to investigate the quality of the recommendations produced by our model.

We introduce TOEFL scores into our model as a topical prevalence covariate. Prior work on student application behaviors shows that the scores students receive on standardized tests influence where they will apply to (Sawyer, 2007). Considering the time and financial costs of putting together an application, students self-select based on their own estimate of the probability of acceptance to that institution. By introducing TOEFL scores as a topical prevalence covariate we aim to explore whether or not TOEFL scores impact the probability that students are members of certain preference-groups and ultimately what institutions they prefer.

We focused our analysis on TOEFL scores reported in 2015 and 2017 and used year as a categorical topical content covariate. These two years allow us to consider the possible impact of the 2016 U.S. presidential election on school preference. International application trends are sensitive to global and national events and the impact those events have on U.S. foreign policy (Laws & Ammigan, 2020; Rose-Redwood & Rose-Redwood, 2017) . In 2016, the rate of international student enrollment decreased by 3% (Institute of International Education, 2020) and this trend has continued through 2020. While there are many potential explanations for this reduction in students (Laws & Ammigan, 2020), it is unclear whether the school preferences of students have also changed in response to these pressures. By introducing the year students reported their scores as a content covariate, we can directly compare how the preferences across institutions changed between these two years.

### 4.1 TOEFL Dataset

Each year, ETS administers the TOEFL test in more than 200 countries and territories (Educational Testing Service, 2021a) with both in-person and at

home testing options. These test-takers vary in age, educational goals, and how they plan on using the test score. Additionally, individuals can take the test multiple times in a given year. For each test taken, test-takers can send up to 4 score reports to institutions for free; additional score reports can be sent for a cost. With the focus of our research on prospective international students (non-U.S.) who are applying to undergraduate programs in the U.S., we took several pre-processing steps prior to analyzing our data. We limited our dataset to score reports associated with TOEFL tests taken in 2015 and 2017. Of these tests, we only considered score reports sent to institutions during the application cycle (e.g. January 2015-June 2016, and January 2017- June 2018). These liberal criteria capture early admissions through late enrollment reporting. Since our focus is on U.S. application decision-making, we did not include reports to non-U.S. institutions in our dataset. All data was anonymous, with no personally identifiable information used in the analysis. Our research plans underwent ethical review by ETS's Committee for Prior Review of Research.

Previous research indicates that graduate and undergraduate students have distinct application behaviors and decision-making (Nicholls, 2018). For this study, we limited our analysis to undergraduates. We considered three factors when determining the education level of an individual: self-reported information, institution and program choice, and age. Of the 5% of test-takers who self-reported educational information, we included students who identified as applying to undergraduate programs. We next removed the students applying to professional and graduate programs from our dataset. The majority of students, however, did not have clear information from either of these sources. There is a clear bimodal distribution in the age of test-takers for many countries. We used a K-means classifier to estimate the most likely breakpoint between the latent age distributions in the data. We labeled as likely undergraduates test-takers younger than 20 years old with no clear self-report or application evidence. We estimated that undergraduates made up approximately 30% of TOEFL takers who sent reports to the U.S.

Not all test takers choose to report their test scores and not every test taker requests the same number of score reports. To reduce some of the sparseness within our dataset we removed institutions that appeared less than four times (811 institutions) and students who only sent applications to a single institution. This produced our final dataset of 113,397 unique individuals who took a total of 130,789 exams, for which a total of 962,618 reports were delivered across 1824 higher education institutions. We included test retakes in our dataset and considered each exam as a unique instance. This means that students who retook tests might appear multiple times in our dataset. We considered each resending of scores as evidence of their application goals, acknowledging that these goals may change between retakes.

### 4.2 Model Fitting

We used the STM package in R (Roberts et al., 2019) to compute the matrix of recommendation scores (i.e., preferences) from test-taker's score reporting behavior and metadata. We explored the use of TOEFL iBT score as a topical prevalence covariate and application year as a topical content covariate. We estimated the impact of TOEFL score on topics using a spline since we expected a nonlinear relationship between score and some groups. We tested models with four distinct covariate structures: 1) including both covariates, 2) only TOEFL score as a topical prevalence covariate, 3) only the year the test was taken as a topical content covariate, and 4) no covariates. The model with no covariates is akin to a CTM (Blei & Lafferty, 2007). This reflects a traditional CF approach, as opposed to a hybrid approach tested by the other three models. For each covariate structure, we fit 11 different models with varying number of topics, from 5 to 105 topics by increments of ten. Each model was initialized using spectral decomposition of the institutions co-occurrence matrix (Roberts et al., 2016). We used STM's built in approximation-based variational expectation-maximization (EM) algorithm to estimate parameters (Roberts et al., 2016) and set the maximum number of EM iterations at 500 (all models converged prior to reaching this threshold).

### 4.3 Model Search

There are numerous ways to evaluate a recommender system and these vary based on the data, algorithm, and goals of the problem. In the current paper we want to assess the ability of our algorithm to 'find good items' for a student who we only have positive choice data for and we expect will only view a limited number of the institutions that exist in the entire dataset. Given these constraints we choose the receiver operating characteristic (ROC) curve based measure, AUC (area under the ROC curve) to support comparison between models of varying complexity (Schafer, Frankowski, Herlocker, & Sen, 2007). AUC measures how well the algorithm can distinguish signal from noise. Our STM model calculates a 'score' for every institution, which can be used to produce a ranked list of institutions for each student. We assume there is a filter-tuning value such that students see all items above a given score. We use AUC to measure our ability at various cutoff points to recall the institutions students actually chose. This metric is not sensitive to the order of items, merely that they occur above that cutoff. A high AUC score would tell us that if we set our cutoff criteria very high, most of the institutions that students sent their score to will appear within the set of items that passed that criteria.

We used a hold-out validation method to evaluate our 44 models to determine how many topics are needed to best capture the latent preference-groups present within our dataset and whether the inclusion of covariates improved predictive performance. For each model, we ran five train-test splits estimating held-out AUC. For each train-test split we select a random subset of students

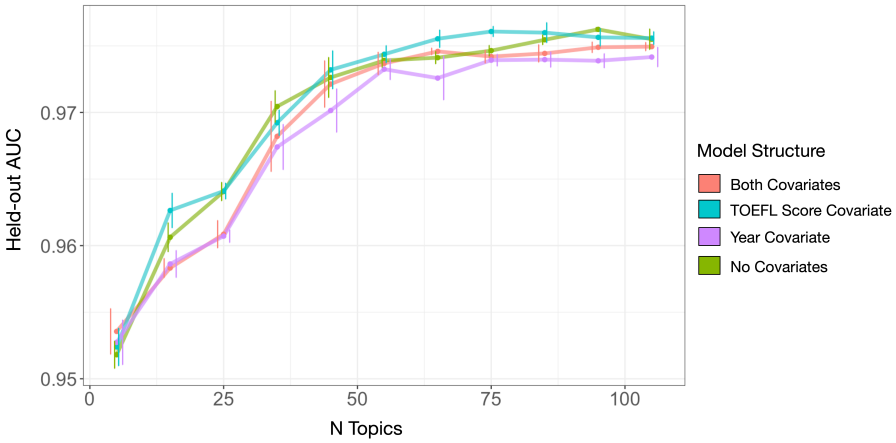


**Fig. 2** Mean AUC for our held-out participants for 5-105 topic models. Error bars reflect 1 standard deviation in the means across 5 hold-out validation folds. Colors reflect the 4 different covariate structures we tested.

(10% of the all students in the dataset) and for each student in that subset we select half of the institutions they applied to at random to hold out from the training data. We calculated AUC for each of our held-out samples and report the averaged values across samples and iterations (Figure 2). Across the 4 covariate structures, we find that AUC improves as we increase the number of topics, steeply increasing before the 65-topic model and leveling off with AUC values around .975.

We want to identify models that not only have good predictive accuracy but also make meaningful discrimination between preference groups. Our two best fitting models are the 75-topic model with TOEFL score as a topical prevalence covariate and our 95-topic model with no covariates. We reviewed the topics for both of these models to determine the extent to which topics varied across models. The 95-topic model provided greater distinction than the 75-topic model, frequently splitting topics in the 75-topic model into smaller groups of institutions. For example, one of our popular topics in the 75-topic model that focused on competitive private R1 institutions (Table 1) was split between two topics within the 95-topic model. One topic favoring University of Pennsylvania and Columbia, and the other favoring on Stanford, Harvard and MIT. Although this is likely picking up on variation in the concurrences between certain institutions, these topics, dominated by only a few institutions, capture preference for specific institutions rather than a type of institution.

We focus the remainder of this paper on the simpler 75-topic model with TOEFL score as a topical prevalence covariate. The average held-out AUC for this model is .976 (SD = .0005). While AUC provides a useful value for model comparison when there is uncertainty about how many items a user is likely to view, it can be difficult to interpret. Precision and recall at K, on the other hand, consider the performance of the model with a specific cut off criteria

K. By setting K to 25, we are assuming that individuals will only consider the top 25 recommendations a model makes. Precision captures the fraction of the top 25 recommendations that were actually chosen by the individual. We find the average precision at 25 of the 75-topic model is .08 (SD= .003). Recall captures what fraction of all items chosen by the individual appear in the top 25 recommendations. The average recall of the 75-topic model is .53 (SD=.02). These values suggest that our model recommends on average half of held-out items within the top 25 recommendations.

Our goal of this paper is two-fold: 1) to model and understand students' latent preferences using STM and 2) to present a case-study for using STM to support students' as they explore institutions to find a good match. To support these aims, the analyses we report in Section 4.4 and Section 4.5 are performed on a 75-topic model we fit to our full dataset. We choose to fit to the full dataset because we are using the model to understand our population of students (similar to a traditional regression framework; Roberts et al. 2014) and are exploring how the model in use would support students in extending beyond the choices they have made, rather than to predict interest in chosen institutions. These results are not meant to provide evidence for the validity of the model predictions.

## 4.4 Results

### *4.4.1 Characterizing Performance Groups*

Out of the 75 topics fit by the STM model with TOEFL score as a topical prevalence covariate, we present 10 topics. The first 5, reflect the institutions associated with the 5 most frequent topics that occurred within our dataset (e.g., the highest probabilistic sum across students) (Table 1). Additionally, we present 5 topics that highlight less frequent but specialized groups, which have varying correlations with other topics (Figure 3). These 5 topics demonstrate how the model is able to capture niches despite their infrequent occurrence within the overall population. For each topic, we present in order the 5 institutions with the highest FREX scores. The FREX metric is calculated for each institution in each topic and balances the overall frequency of an institution in the dataset with its exclusivity to that topic (Bischof & Airoldi, 2012; Roberts et al., 2019). We used the Carnegie Classifications and U.S. News Rankings of the institutions most strongly associated with each topic to characterize these topics (Indiana University Center for Postsecondary Research, n.d.). These characterizations do not necessarily apply to all institutions within a group nor do they reflect the characteristics students are aware of when selecting these institutions. These labels are intended to capture the type of institution associated with each preference group.

The STM model estimates the probability that each institution is in each topic (e.g., the probability the institution is selected by that preference-group). This means it is possible that some institutions are likely to appear across

**Table 1** Characterization of 10 select topics using descriptions from the Carnegie classification and IPEDs (Indiana University Center for Postsecondary Research, n.d.; National Center for Education Statistics, 2021)

| Label | Characteristics | Associated Institutions (ranked by FREX score) |
|---|---|---|
| Land-Grant | Large public R1 institutions where bachelor's degree majors are equally balanced between arts and sciences, and professional fields. | Univ. Connecticut, Ohio State, Indiana Univ. (Bloomington), Pennsylvania State, Univ. of Pittsburgh |
| Private R1 | Private, highly selective R1 institutions. The majority of students enrolled at these institutions are graduate students. | Stanford Univ., Massachusetts Inst. of Tech., Univ. of Pennsylvania, Harvard, Columbia |
| Tech. Focus | Primarily large public R1 institutions that are well known for strong science and engineering programs. | Univ. of Illinois (Urbana-Champaign), Georgia Tech., Univ. of Wisconsin (Madison), Univ. of Michigan (Ann Arbor), Rosehulman Inst. of Tech. |
| Public R1 | Large public R1 institutions where 60-79% of bachelor's degree majors were in arts and sciences. | Univ. of Virginia, Univ. of North Carolina (Chapel Hill), Emory Univ.,Washington Univ. (St. Louis), Univ. of Michigan (Ann Arbor) |
| U Cal. | Institutions within the University of California (UC) system. | UC San Diego, UC Santa Barbara, UC Davis, UC Irvine, UC Santa Cruz |
| Cal. State | Institutions within the California State (CS) University system. | CS Long Beach, CS Los Angeles, San Francisco State, San Jose State, CS Northridge |
| CA Comm. | Community Colleges located in Northern California. | De Anza Col., Foothill Col., Diablo Valley Col., Ohlone Col., Col. of San Mateo |
| Arts | Institutions focused on Art and Design. | Art Inst. Chicago, Maryland Inst. Col. of Art, Sch. of Visual Arts, Rhode Island Sch. of Design, Ringling Col. of Art and Design |
| Music | Music Conservatories (Cons.) and institutions with well known music programs. | Juilliard, New England Cons. of Music, Manhattan Sch. of Music, San Franscisco Cons. of Music, Cleveland Inst. of Music |
| SLACs | Highly selective small Liberal Arts colleges (SLACs) and universities. | Gettysburg College, Connecticut Col., DePauw Univ., Trinity Col., Skidmore Col., |

multiple topics. We find that cross-topic listing is most prevalent for some highly popular institutions (e.g. Purdue, Penn. State, Ohio State). In Figure 3, we present a network graph showing the correlation between topics where the size of the node labels represents the scaled proportion of those topics within our dataset (larger text for more frequent topics) and the edge width between topics shows the pairwise topic correlation. We labeled the topics in Table 1 and numbered the remaining topics.

The large cluster at the lower left of the network plot demonstrates the high correlation between many of the topics capturing many of the different

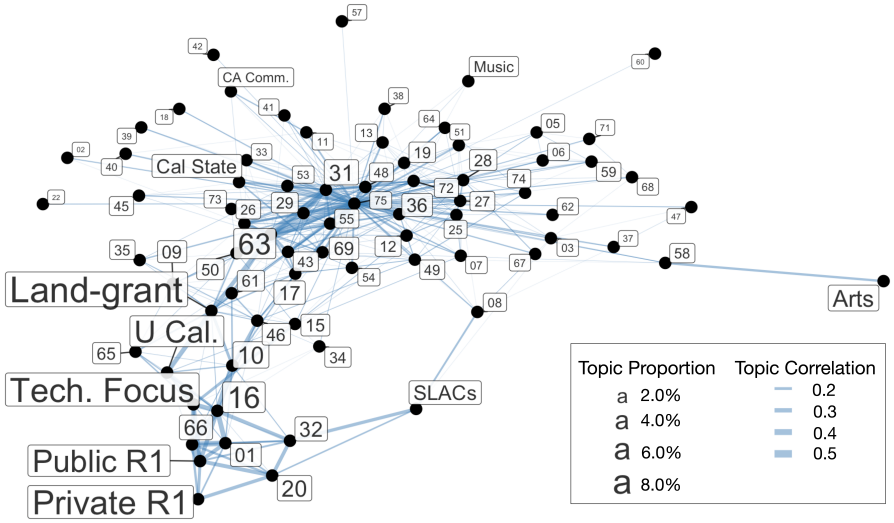


**Fig. 3** Network plot of the latent preference-groups. Text size indicates the proportion of topic within the dataset, edge size indicates the correlation of topic-institution distributions between topics. Nodes characterized in Table 1 are labeled. We only show edges with correlations greater than 0.2.

dimensions present in large research institutions (e.g. public, private, selective, degree specializations). This figure illustrates how some of the more specialized topics, such as the Art and Music schools are much less correlated with other topics. These specialized preferences contribute to the hub-and-spoke network structure at the center of Figure 3. At the center we see topics dominated by large state schools (e.g. topics 75, 31, 28, 19) and private universities (e.g. topics 43, 36) that have high acceptance rates and offer a wide variety of programs. On the fringes we see institutions that appeal to more niche areas of study.

We found many topics that captured regional preferences for places such as California, Florida, Texas, and Massachusetts (e.g. topics 69, 74, 12 and 17 respectively). These regional preferences appeared across institution type, size, and competitiveness. For example we see distinct preferences for California institutions that separate the University of California system, from California State Colleges, and Community Colleges (Table 1). Within California Community Colleges we even see a distinction between Northern and Southern California (CA Comm. and topic 42).

### *4.4.2 TOEFL Score as a Prevalence Covariate*

We used STM to perform the matrix factorization for our hybrid CF approach because we hypothesized that test-takers' TOEFL scores would influence their choices and college preferences, and thus the topic membership. In Figure 4,

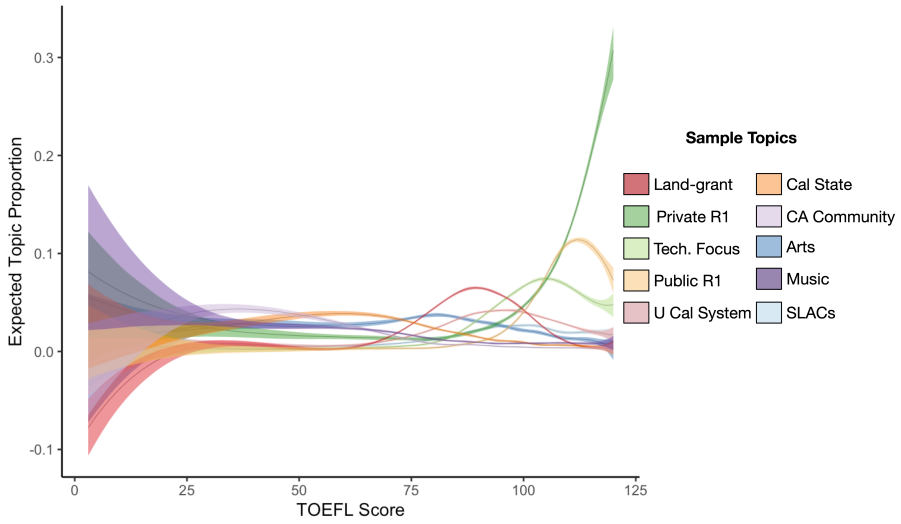


**Fig. 4** Test-taker's TOEFL scores affect the expected proportion a given topic is represented by a test-taker. We show the expected topic proportions for topics characterized in Table 1. Topics are plotted as a smooth function of TOEFL score with shading surrounding line representing 95% confidence intervals.

we plot the relationship between TOEFL score and the expected proportion of a student with that score that is captured by the topics characterized in Table 1. TOEFL scores appear to impact which preference groups they belong to; however, this effect is not uniform across groups. Our Private R1 Institutions (dark green) and Public R1 Institutions (light orange) topics are largely associated with students with high TOEFL scores. In contrast, we find that our Music (dark purple) and CA Community College (light purple) topics are associated with students' TOEFL scores below 80. This relationship is not necessarily linear, and we see this in topics such as Land-grant universities (red), a topic that is most associated with students who score around 90 on TOEFL.

### *4.4.3 Year as a Content Covariate*

Our comparison between different models found that the addition of year as a content covariate did not improve held-out AUC. To explore the inclusion of year within these models of preference, we consider the 75 topic model that used the year students took the TOEFL as a topical content covariate and score as a topical prevalence covariate. This allows us to look at how the same topic changes between test years. For each year we calculated the difference in $\beta_{d,k}$ between the same topic across the two years. The topic-institution distributions, $\beta$, captures the probability that institutions are associated with a given topic (see Section 3).

Figure 5 shows the impact of year on our two most frequently occurring topics (Land-grant universities and Private R1s). This figure shows which institutions experience change in this topic across year. Institutions further to the left (scaled $\beta_{d,k}$ closer to 0) are more representative of the topic in 2015, and institutions further to the right (scaled $\beta_{d,k}$ closer to 1) are more representative of the topic in 2017. Institutions not shown and institutions close to the center (scaled $\beta_{d,k}$ closer to .5) are equally representative of the topic regardless of the year. For both of these topics, the changes in the topic-school probability is highly correlated with the changes in the number of scores sent to these institutions between the two years (Private R1: r(8)=.86, p<.001 , Land-grant: r(9)=.66, p<.02). Most of the institutions in these topics see a reduction in international student interest between these two years; the model interprets the size of the reduction as indicative of changes in how students express their preference. In our earlier consideration of the 95-topic model with no covariates in Section 4.3, we saw our model split the Private R1 topic, creating a separate group for University of Pennsylvania and Columbia. Exploring the influence of year as a topical content covariate suggests that the added complexity of the 95-topic model may be fitting variation across year by creating separate topics to compensate for changes in the relative popularity of schools in the same topic.

## 4.5 Implications for Recommendation

### *4.5.1 Beyond-Accuracy Evaluation*

Beyond-accuracy evaluations shape the user experience by capturing different properties of the recommendations generated by a recommender system (Kaminskas & Bridge, 2016; Shani & Gunawardana, 2011). Unlike accuracy-based evaluation (Figure 2), these beyond-accuracy measures aim to quantify behavior of the recommendations the system across test-takers opposed to evaluating whether the recommendations align with test-takers' choices. We ran several beyond accuracy evaluations for our final 75-topic model that included only TOEFL score as a prevalence covariate.

We first consider the item-space coverage and spread of the top 25 recommendations our system generated for all test takers. Coverage captures the aggregate number of distinct items recommended to users as:

$$Coverage = \mid \cup_{u \in U} R_u \mid \tag{6}$$

where $R_u$ is the set of top-N recommendations generated for user $u$ and $U$ is the set of all users. In Figure 6a, we graph the percent of all institutions that are recommended across the top-N institutions (1-25). When we look at the most probable recommended institution (N = 1) across all test-takers we find we only cover a small percent of the space of institutions (e.g. 75 institutions, 4.1% of 1824 institutions within our sample). Coverage increases as we expand the number of top items our system recommends institutions. More than half

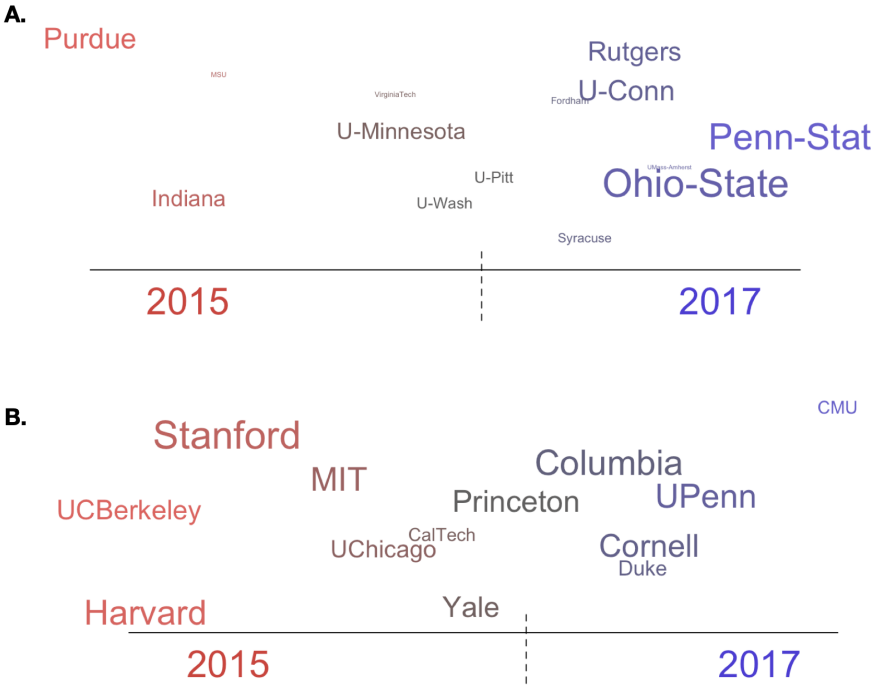


**Fig. 5** A) Institution inclusion across years within the Land-Grant Topic and B) Institutions within the Private R1 University Topic. Distance from the central dotted line indicates the scaled-change in topic-institution distribution between 2015 and 2017. Font size is proportional to the institution's occurrence within the data set, with institutions such as Ohio State occurring more frequently than Syracuse.

of the institutions in our data set appear in the top 25 recommendations for at least one student.

Spread captures how the recommender system spreads attention across all items (Shani & Gunawardana, 2011). We calculate this as the normalized entropy of institutions recommended to students:

$$Spread = -\sum_{i=1}^{n} p(i)log(p(i))/log(n) \tag{7}$$

where each institution $i$ accounts for a proportion $p(i)$ of all recommended institutions. We normalize Shannon Entropy between 0 and 1 by dividing by $log(n)$ which captures the maximum possible diversity index across all $n$ items. This means the closer the value is to 0 the more similar the recommendations are across different users. On the other hand, a system that provides unique recommendations to each user would have a score of 1. We find that our recommendations are least diverse when we are looking at the top few items, but the set of recommendations provided to students becomes more diverse and plateaus as we increase the number of recommendations shown to a student

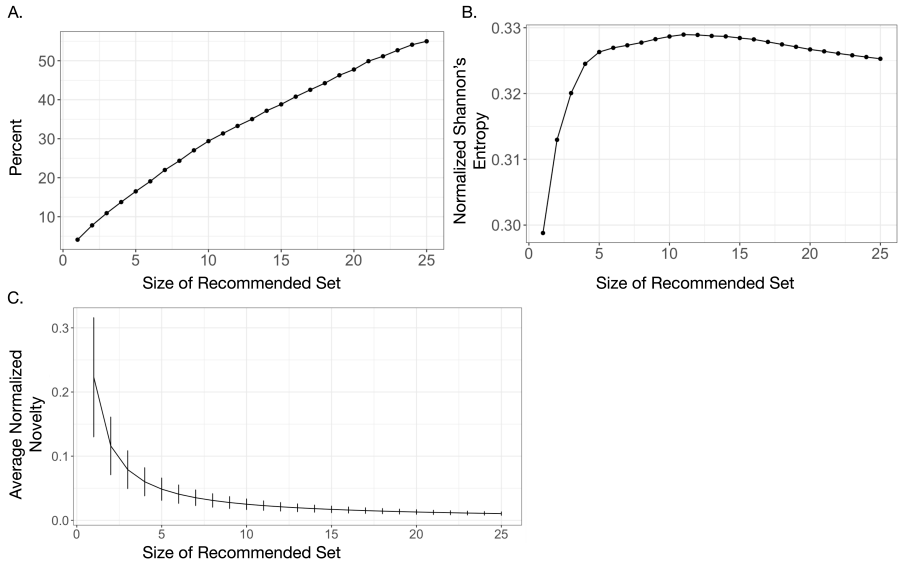


**Fig. 6** Item-based beyond-accuracy measures A) Coverage: percent of institutions in entire dataset recommended across top 1-25 recommendations. B) Spread: Entropy of recommendations across top 1-25 recommendations.C) Novelty: The average normalized novelty of institutions within the first 1-25 recommended institutions.

(Figure 6b). The baseline spread across all student choices in our dataset was 0.38, indicating that our recommender is providing recommendations that would increase similarity across student choice.

Finally, we consider the user-specific metric of novelty. Novelty measures the ability of the system to recommend uncommon items to a user (Figure 6c). We applied the novelty calculation proposed by Kaminskas and Bridge (2016) to consider how novelty changes as we increase the number of recommendations $R$ we provide a student:

$$Novelty(R) = \frac{1}{nov_{max} \cdot \mid R \mid} \sum_{i \in R} -log \frac{\mid u \in U, r_{ui} \neq \emptyset \mid}{\mid U \mid} \tag{8}$$

We calculated novelty as the log of the average popularity of items recommended to a user $u$, where $U$ is the set of all users in the dataset. We further normalized this value by dividing the average novelty by $nov_{max} = -log(\frac{1}{|U|})$, the maximal possible novelty value for a set of recommendations. The closer this value is to 0 the the greater the overall popularity of the institutions recommended to a given student. The baseline novelty across all student choices in the dataset we trained our model on was quite low at 0.04 (SD=.05). We find that novelty is highest for our first 5 recommendations; however, as we increase the number of recommendations the average novelty decreases. The higher novelty for the top items is driven by our niche topics such as our CA Community Colleges and Arts topics, which feature a diverse set of infre-

quently selected institutions. However, we find that expanding the number of items selected leads the recommender system to increasingly pull from generally popular items.

### *4.5.2 Recommendations for Simulated Students*

We created several simulated students to demonstrate how different interests are treated by our recommender when students have different TOEFL scores and tested in different years. We represent 3 different preference types and scenarios. For each scenario we used our model to estimate the preference of two students who are identical in their selected institutions but differ in their TOEFL scores (Table 2).

For our first student we select three institutions: one in the University of California (UC) system, one in the California State system, and one in the California Community College system. This student represents someone who is focused on studying in California but is exploring the different systems or is unaware of the distinctions among these institutions. We find that our recommender favors California Community and State colleges over UCs when the student's TOEFL score is 70. When their score is 100 the preference shifts towards UC schools but also expands to selective institutions outside the state. Our second student selects three competitive liberal arts schools from a Forbes list published in 2016 specifically highlighting 'best' liberal arts colleges for international students (Wang, 2016). For this student, the recommendations produced by our system are the same despite differences in the TOEFL score. The institutions recommended consist of liberal arts and private universities on the East Coast. Finally, our third scenario considers how the inclusion of TOEFL score within the model can improve the focus of recommendations when only one institutions of interest is selected. This captures how the system behaves with very little information. For this student we see that a TOEFL score of 70 leads the model to suggest institutions known for their music programs and institutions with higher acceptance rates. On the other hand, when students have a higher TOEFL score the model suggests popular selective private and public universities.

## 5 Discussion

In this paper, we propose a 'student-centered' approach that uses students' choices and meta-data to support personalized recommendations to support the student-institution matching process. We present a use-case for Structural Topic Modeling (STM) as a novel approach for supporting recommendations that pairs student choices with the metadata surrounding those choices. By applying this approach to test-takers' TOEFL score reporting behavior, we are able to explore their latent preferences. We test hypotheses concerning how those latent preferences relate to TOEFL scores and vary across years and thus, produce recommendations that are sensitive to both information

**Table 2** Top 5 recommendations for simulated students with varying TOEFL scores

| Student Selection | Rank | Recommended Institutions | |
|---|---|---|---|
| | | TOEFL: 70 | TOEFL: 100 |
| Cal. State Fresno | 1 | San Jose State | U. Washington (Seattle) |
| UC San Diego | 2 | De Anza Col. | UC San Diego |
| De Anza Col. | 3 | Santa Monica Col. | UC Irvine |
| | 4 | Cal. State (Long Beach) | UC Davis |
| | 5 | San Diego State | Boston Univ. |
| Swarthmore Col. | 1 | Tufts Univ. | Tufts Univ. |
| Williams Col. | 2 | Dartmouth Col. | Dartmouth Col. |
| Wellesley Col. | 3 | Wesleyan Univ. | Wesleyan Univ. |
| | 4 | Brown Univ. | Brown Univ. |
| | 5 | Colby Col. | Colby Col. |
| Johns Hopkins | 1 | Univ. of Cincinnati | Boston Univ. |
| Univ. | 2 | Oberlin Col. | Cornell Univ. |
| | 3 | Michigan State | UC Berkeley. |
| | 4 | Arizona State | Univ. Pennsylvania |
| | 5 | Johns Hopkins Univ. | Univ. Illinois (Urbana-Champaign) |

about the test-taker and the context of their choice. Finally, we use beyond accuracy measures and simulated students to explore how this model would behave within a Recommender System.

### 5.1 STM as a Model of Latent Preference

We present results from the STM model that identified 75 latent topics present within our data set. These topics distinguished preference-groups that differed in terms of program focus, institution size, type of institution, geography, and acceptance rates. We found that a test-taker's TOEFL score had a widely variable impact on the predicted probability that test-takers were members of specific preference-groups. Test-takers with high TOEFL scores had a greater probability of belonging to groups characterized by institutions with lower acceptance rates (in other words, more selective institutions), than students with lower scores. While test-takers are free to apply to any institution, their score-sending behavior is sensitive to the relative standing of their test scores against the score requirements of each school. This study was limited to exploring TOEFL scores and preference choices; however, future inclusion of variables that reflect students' academic performance (e.g. High School GPA, ACT/SAT scores), as well as other types of cognitive and non-cognitive skills, could allow us to capture more variation in test-taker preference. This work may be especially significant for identifying recommendations for competitive institutions where TOEFL scores are primarily used as a baseline criterion of English language skills.

We fit a model with year as a topical content covariate to test whether student preference varied by the year they sent their scores. Exploring the

topics the model estimated, we found some variability in how representative certain institutions were of the preference group. This variation was significantly correlated with changes to total reporting volumes received by those institutions. Additionally, we found evidence that our model without covariates was dividing topics to capture these changes between years. While our model including year as a covariate did not perform as well as simpler models as measured by held-out AUC, the descriptive value and ability of the topics to adjust for fluctuations in volume suggest that the inclusion of this covariate may be merited. While STM provides a useful tool for testing the relationship between covariates and latent preference groups, it does not allow us to make causal claims. In this study, we could not test whether the changes between 2015 and 2017 were due to large-scale geopolitical trends or caused by smaller, more targeted events. However, when specific events occur for a given institution, we can analyze the change in topic loadings to investigate the impact of that change on student preferences. In future research, we can explore the usefulness of this covariate structure in capturing the effect that changes in individual institutions' admission policies or marketing campaigns have on topic-institution distribution. Additionally, as we gather data about the effect of the COVID-19 pandemic on student application behaviors we can use year to understand whether the pandemic has impacted preference and how.

We chose TOEFL score and test year as topical prevalence and content covariates because the test-takers records were readily available and prior research suggests that these are important factors. Our findings should encourage expansion of the factors, and for this purpose, we have demonstrated that STM provides a useful tool for not only incorporating this additional information into models of student preference, but also understanding the relationship between these factors and preference. Finally, it should be noted, there are limitations to how many factors the STM approach can incorporate. The approach can support multiple topical prevalence covariates and account for interaction effects; however, STM can only support content covariates with a few levels and in these cases, the model requires more data and can be slow to converge. Ultimately, the appropriateness of different recommendation algorithms varies depending on the data available, the framing of the problem, and the type of Recommendation System these models will be embedded within (Çano & Morisio, 2017). While STM was well suited for our data, future work extending the metadata considered within the models should explore other algorithms that may be better able to take advantage of the data available.

## 5.2 STM to Support Recommendation

### *5.2.1 Qualitative Evaluation*

Our exploration of the model's latent preference-groups shows the impact of metadata on the likelihood test-takers belong to those groups and how those groups' preferences change across time. To understand how metadata impacts

the type of recommendations our system makes, we simulated several different test-takers while controlling for their TOEFL scores. Overall, we found that our recommender produced recommendations that were responsive to the interests our simulated test-takers expressed for different regions and institution types; however, the impact of metadata on these recommendations varied based on the test-takers' expressed preferences.

We find that TOEFL score has a much greater impact on our recommendations for students that select a set of institutions with widely differing acceptance rates and minimum TOEFL requirements. In these situations the covariate structure estimated by the model identifies institutions that are more likely to appeal to the student given their score. On the other hand, when a student targets a particular type of school (in our example SLACs) this produces a strong signal of preference which variation in TOEFL score does not alter the recommendations the model produces. Our final simulation provides an example of how the model performs with only a single choice. In the case of Johns Hopkins, we have a school with different minimum TOEFL requirements for different areas of study. With lower TOEFL requirements for music students, the model recommends institutions well known for their music programs (University of Cincinnati and Oberlin) and higher acceptance rates (MSU and ASU) to the student with a TOEFL score of 70. On the other hand, when the student has a score of 100, recommendations favor private and public institutions with similar competitive undergraduate admissions.

Unlike rule-based approaches that can modify recommendations based on whether students meet university requirements (Ragab et al., 2012), our approach only captures the implicit impact of score on preference. Not only does this make our approach easier to maintain, it also means our model can be sensitive to variation in how institutions use TOEFL as part of the admissions process (e.g. as a requirement or an option, as a universal standard or varying between programs). From the perspective of system design and fairness, it is unclear how best to account for students' scores in generating recommendations. The TOEFL test is meant to primarily capture student's readiness to study at English-medium institutions (Educational Testing Service, 2021b) and does not capture the full qualifications of an applicant. While score may provide information to help identify student preferences, future research will need to investigate whether the inclusion of this covariate introduces biases into the model which overly limit students' recommendations. This future work would benefit understanding how students decide to apply to 'reach' schools, when to suggest institutions outside of a student's perceived qualifications, and what support would increase their chance of acceptance and success at that institution.

#### *5.2.2 Quantitative Evaluation*

There is no single metric for determining what makes a good recommender; instead, researchers must rely on a variety of measures to consider how the system meets the different preferences, intentions and needs of the system's users

(Schafer et al., 2007). We therefore ran several offline evaluations using both accuracy-based and beyond-accuracy-based measures to better understand the behavior of our system. These measures helped us identify how future research and online evaluation could improve how we support college choice.

*Student Preference.* Accuracy-based metrics and similar quantitative measures that compare predicted and true ratings are one measure of recommendation quality (Schafer et al., 2007). These metrics tell us whether the recommendations our system provides are sensitive to the observed preferences of the student. These methods can be challenging to use with sparse datasets, especially when using positive only data rather than rankings. When data is sparse, many accuracy-based methods will introduce bias in the models that describe users who make large numbers of choices and recommendations favoring the most popular choices. Using hold-out validation, we see sharp improvements to AUC that levels off around 75-topics. We find little difference in the accuracy of the recommendations generated by our four models with different covariate structures. Future work would benefit from exploring additional measures of student preference that could be used to compare these different models. Laboratory studies, while limited in their external validity, can provide us with user-feedback on what drives student preferences and on how well system recommendations align with those personal preferences. This will be helpful in improving both our system's detection of niche preferences and disambiguating the preferences that drive selection of very popular institutions.

*Student Intention.* A student's intentions for using our recommender system should shape how we evaluate our system. A student beginning to explore institutions will likely appreciate different recommendation behavior from a student who is looking to add a few more institutions to their top choices. Using beyond-accuracy measures, we could explore the different properties of the recommendations generated by our system. Measures such as coverage, spread, and popularity, capture the range of items that the system can make recommendations about as well as how the system spreads attention across that set of items. A system with low coverage will produce similar recommendations across individuals. We find that our system has high coverage, with the top 25 recommended items capturing over 50% of institutions. High-coverage is ideal if our system will support matching across a wide range of institutions in the US. The spread of recommended items levels off after the first 5 recommendations as we increase the number of recommendations the system provides. Popular institutions, such as University of Michigan, feature prominently across many students despite different preferences and TOEFL scores. There is often a trade-off between the spread of the recommendations generated and accuracy of the model. With 72% of test-takers selecting at least one of the 25 most frequently chosen schools, a system sampling from these institutions is not only likely to improve held-out AUC, precision, and recall, but would also decreases the spread of the recommendations. To some

extent, we want popularity to influence recommendation as these institutions are popular with international students for many valid reasons (e.g. availability of pathway programs, international name recognition, strong support communities, diverse program offerings). While this behavior does not support users looking for lesser-known institutions, the trend we see in our coverage measure suggests that as we provide more recommendations these lesser-known institutions enter the space of recommendation. Future iterations of this system will need to survey potential users to identify the right balance of coverage and spread to support user preference.

Measures such as novelty, serendipity and diversity are useful when the goal of the system is to generate relevant recommendations that are unknown to the user, surprising, or different from what they are considering. We find that novelty is greatest for our top recommendations, and increasing the number of recommendations our system provides decreases the average novelty of the recommended set. This is not surprising given the high correlation among many of the most frequent topics seen in Figure 3. One way in which we can explore increasing novelty in future work is by applying a secondary weighting criterion that penalizes more popular institutions. Increasing novelty can expand the number of institutions a student considers and aid in expanding preferences. When the preferences of the student are driven by rigid constraints, however, this behavior can seem random and be off-putting. As with other beyond-accuracy measures, future work will need to survey students at different stages of the application process to identify what kind of system behavior best complements the intentions of the student.

*Student Need.* Student needs change across the stages of the application process. Our score reporting data provides a limited view of the multistage process of matching prospective students with institutions. Collecting data from students over the course of this process faces numerous challenges; however, with this information it would be interesting to examine the relationship between our models of preference and the institutions students eventually apply to, are accepted at, and then choose to attend. This information could provide a comparative measure for optimizing system performance or be used directly as a means of further re-ranking recommender outputs according to a student's probability of matching a given institution.

What qualifies as a 'good match' between a student and a university can also be captured in outcomes that follow the admissions process. While there are many issues with predicting academic, social-cultural, psycho-social, and career outcomes (e.g. Mesidor & Sly, 2016), these models have potential value for supporting at-risk students. Throughout this paper, we have described our system as modeling student preference; however, it is unclear whether students' preferences are actually predictive of student success at a given institution. With more complete information about a student's academic trajectory, it would be valuable to explore the relationship between our models of preference and these different dimensions of success.

## 6 Conclusion

Student-Institution matching is a complex problem that has been explored from numerous angles. Approaches from the field of economics have been influential in improving the fairness of centralized application processes, but are limited by their ability to model a system in which the preferences are often only partially known (Che & Koh, 2016). Predictive models of success, while useful for identifying at-risk students, are problematic when used for admissions (Alyahyan & Düştegör, 2020; Holmes et al., 2021; Hutt, Gardener, Kamentz, Duckworth, & D'Mello, 2018). In this paper we explore hybrid collaborative filtering as an alternative, 'student-centered' approach to support matching. By modeling students' preferences and identifying institutions that best align with those preferences we aim to improve matching by supporting students' exploration of institutions and reducing the likelihood that good matches are not considered. We present a case study for using Structural Topic Modeling (STM) as a means for modeling how student factors and the context of their choices impact their preferences for schools. We demonstrated how STM can be used to understand the influence of these factors (i.e. TOEFL score and year) on the expression of student preference. Unlike other hybrid approaches that can be challenging to interpret and make sense of (Çano & Morisio, 2017), STM provides insight into the contribution of different factors, making it well suited for supporting both fundamental research on student preference and the design of an operational college recommendation system.

Prior approaches for recommending institutions to students (e.g. Bokde et al., 2015; Iyengar et al., 2017; Ragab et al., 2012) have used methods that require standardized features across students and institutions and extensive data about students and their choices. These requirements limit how well these approaches can scale to support making recommendations of a large number of institutions for a diverse set of students. In this paper, we demonstrate that the choices students make contain valuable information from which we can infer their preferences. While the TOEFL dataset we described is unique to ETS, we believe that there are many datasets that would benefit from a similar problem formulation. Within the U.S., datasets from the Common Application (Freeman, Magouirk, & Kaijkawa, 2021), National Student Clearninghouse (Dundar & Shapiro, 2016), and the National Resident Matching Program (National Resident Matching Program, 2021), could benefit from the application of an approach similar to ours. Internationally, countries that use centralized admissions systems (e.g., Germany, Taiwan, Turkey, Chile), collect data that could also benefit from this type of modeling approach. While previously published work has used these datasets to predict student success (Hutt et al., 2018) and to build and test matching algorithms (Braun, Dwenger, Kübler, & Westkamp, 2014; Westkamp, 2013), using such data to understand student preference would provide a complementary and important perspective on the student-institution matching problem.

In summary, the findings we present in this paper contribute to the areas of research on Student-Institution Matching and Educational Recommender

Systems (Deschênes, 2020; Rivera, Tapia-Leon, & Lujan-Mora, 2018). Future work should focus on exploring additional student and context-relevant features within the model, and testing how the model recommendations relate to different student outcome measures. We plan to extend this research through online evaluation of this approach and further investigation of students' reasoning and decision-making as they select institutions.

**Acknowledgements** This work was funded by Educational Testing Service. We would like to thank Madeleine Keehner and Christopher MacLellan for their comments and feedback on drafts of this manuscript.

## Statements and Declarations

This work was funded by Educational Testing Service (ETS). All authors are current employees of ETS.

## References

Abunawas, M. (2014). *A meta-analytic investigation of the predictive validity of the test of english as a foreign language (toefl) scores on gpa* (Unpublished doctoral dissertation).

Alfattal, E. (2016). A new conceptual model for understanding international students' college needs. *Journal of International Students*, *6*(4), 920–932.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21.

Arbeit, C. A., & Warren, J. R. (2013). Labor market penalties for foreign degrees among college educated immigrants. *Social science research*, *42*(3), 852–871.

Bastedo, M. (2021). Holistic admissions as a global phenomenon. In *Higher education in the next decade* (pp. 91–114). Brill.

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th international conference on machine learning (icml-12)* (pp. 201–208).

Black, S. E., Cortes, K. E., & Lincove, J. A. (2015). Academic undermatching of high-achieving minority students: Evidence from race-neutral and holistic admissions policies. *American Economic Review*, *105*(5), 604–10.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bokde, D. K., Girase, S., & Mukhopadhyay, D. (2015). An approach to a university recommendation by multi-criteria collaborative filtering and

dimensionality reduction techniques. In *2015 ieee international symposium on nanoelectronic and information systems* (pp. 231–236).

Braun, S., Dwenger, N., Kübler, D., & Westkamp, A. (2014). Implementing quotas in university admissions: An experimental analysis. *Games and Economic Behavior*, *85*, 232–251.

Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, *21*(6), 1487–1524.

Che, Y.-K., & Koh, Y. (2016). Decentralized college admissions. *Journal of Political Economy*, *124*(5), 1295–1338.

Chen, L.-H. (2008). Internationalization or international marketing? two frameworks for understanding international students' choice of canadian universities. *Journal of Marketing for Higher Education*, *18*(1), 1–33.

Deschênes, M. (2020). Recommender systems to support learners' agency in a learning context: a systematic review. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–23.

Dundar, A., & Shapiro, D. (2016). The national student clearinghouse as an integral part of the national postsecondary data infrastructure. *National Student Clearinghouse Research Center*, *20*, 1–18.

Educational Testing Service. (2021a). *The toefl family of assessments.* Retrieved 2021-11-02, from `https://www.ets.org/toefl/`

Educational Testing Service. (2021b). *Validity evidence supporting the interpretation and use of toefl ibt scores.* (Vol. 4). Retrieved 2021-11-02, from `https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf`

Flaitz, J., Eckstein, L. K., Kalaydjian, K. S., Miranda, A., Mitchell, D. A., Mohamed, A., ... Zollner, E. L. (2003). *Understanding your international students: An educational, cultural, and linguistic guide.* University of Michigan Press Ann Arbor, MI.

Freeman, M., Magouirk, P., & Kaijkawa, T. (2021). *Common app research breifs.* Retrieved 2022-05-04, from `https://www.commonapp.org/about/reports-and-insights`

Guo, M., Zhang, Y., & Ye, J. (2019). Does a foreign degree pay? the return to foreign education in china. *Review of Development Economics*, *23*(1), 415–434.

Hallak, J., & Poisson, M. (2007). *Corrupt schools, corrupt universities: What can be done?* International Institute for Education Planning Paris.

Hemsley-Brown, J., & Oplatka, I. (2015). University choice: what do we know, what don't we know and what do we still need to find out? *International Journal of Educational Management*.

Heuser, B. L., Martindale, A. E., & Lazo, D. J. (2016). Strategic internationalization in higher education: Contexts, organizations, and implications for academic integrity. *Handbook of academic integrity*, 347–364.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... Koedinger, K. (2021). Ethics of ai in education: towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 1–23.

Hossler, D., Chung, E., Kwon, J., Lucido, J., Bowman, N., & Bastedo, M.

(2019). A study of the use of nonacademic factors in holistic undergraduate admissions reviews. *The Journal of Higher Education*, *90*(6), 833–859.

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). Prospectively predicting 4-year college graduation from student applications. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 280–289).

Indiana University Center for Postsecondary Research. (n.d.). *The carnegie classification of institutions of higher education.* (Vol. 2018 edition). Retrieved 2021-11-02, from `https://carnegieclassifications.iu.edu/`

Institute of International Education. (2020). *Open doors report on international educational exchange.* Retrieved 2021-09-30, from `http://www.iie.org/opendoors`

Institute of International Education. (2021). *Economic impact of international students.* Retrieved 2021-09-30, from `https://www.iie.org/Research-and-Insights/Open-Doors/Economic-Impact-of-International-Students`

Iyengar, M., Sarkar, A., & Singh, S. (2017). A collaborative filtering based model for recommending graduate schools. In *2017 7th international conference on modeling, simulation, and applied optimization (icmsao)* (pp. 1–5).

Kaminskas, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(1), 1–42.

Laws, K., & Ammigan, R. (2020). International students in the trump era: A narrative view. *Journal of International Students*, *10*(3), xviii–xxii.

Mazzarol, T., & Soutar, G. N. (2002). "push-pull" factors influencing international student destination choice. *International Journal of Educational Management*.

Mesidor, J. K., & Sly, K. F. (2016). Factors that contribute to the adjustment of international students. *Journal of international students*, *6*(1), 262–282.

National Center for Education Statistics. (2021). *Ipeds : Integrated postsecondary education data system.* (Vol. 4). Retrieved 2021-11-02, from `https://nces.ed.gov/ipeds/`

National Resident Matching Program. (2021). *National resident matching program, results and data: 2021 main residency match.* Washington, DC.

Nicholls, S. (2018). Influences on international student choice of study destination: Evidence from the united states. *Journal of International Students*, *8*(2), 597–622.

Posselt, J. R., & Grodsky, E. (2017). Graduate education and social stratification. *Annual review of sociology*, *43*, 353–378.

Ragab, A. H. M., Mashat, A. F. S., & Khedra, A. M. (2012). Hrspca: Hy-

brid recommender system for predicting college admission. In *2012 12th international conference on intelligent systems design and applications (isda)* (pp. 107–113).

Rivera, A. C., Tapia-Leon, M., & Lujan-Mora, S. (2018). Recommendation systems in education: A systematic mapping study. In *International conference on information technology & systems* (pp. 937–947).

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software*, *91*(1), 1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064–1082.

Rose-Redwood, C., & Rose-Redwood, R. (2017). Rethinking the politics of the international student experience in the age of trump. *Journal of International Students*, *7*(3), I–IX.

Sawyer, R. (2007). Indicators of usefulness of test scores. *Applied Measurement in Education*, *20*(3), 255–271.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291–324). Springer.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257–297). Springer.

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353–382). Springer.

UNESCO. (2018). *Global education monitoring report 2019: Migration, displacement and education – building bridges, not walls.* Retrieved 2021-09-30, from `https://en.unesco.org/gem-report/report/2019/migration`

Verstrepen, K., Bhaduriy, K., Cule, B., & Goethals, B. (2017). Collaborative filtering for binary, positiveonly data. *ACM SIGKDD Explorations Newsletter*, *19*(1), 1–21.

Wang, Y. (2016). *12 best liberal arts colleges for international students: Why say no to stem.* Retrieved 2021-09-30, from `https://www.forbes.com/sites/yizhuwang/2016/09/28/liberal-arts-colleges-are-fast-becoming-a-magnet-for-international-students/?sh=22d6585c4798`

Westkamp, A. (2013). An analysis of the german university admissions system. *Economic Theory*, *53*(3), 561–589.

Wiers-Jenssen, J., & Try, S. (2005). Labour market outcomes of higher education undertaken abroad. *Studies in Higher Education*, *30*(6), 681–705.

Wu, T., & Naidoo, V. (2016). The role of international marketing in higher education. In *International marketing of higher education* (pp. 3–9). Springer.

Xie, W., Dong, Q., & Gao, H. (2014). A probabilistic recommendation method inspired by latent dirichlet allocation model. *Mathematical Problems in Engineering*, *2014*.

Zhu, L., & Reeves, P. (2019). Chinese students' decisions to undertake postgraduate study overseas. *International Journal of Educational Management*.