International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-021-00271-1

ARTICLE

Equality of Learning Opportunity via Individual Fairness in Personalized Recommendations

Mirko Marras[1] · **Ludovico Boratto**[2] · **Guilherme Ramos**[3] · **Gianni Fenu**[2]

Accepted: 26 July 2021Published online: 08 October 2021

Abstract

Online education platforms play an increasingly important role in mediating the success of individuals' careers. Therefore, while building overlying content recommendation services, it becomes essential to guarantee that learners are provided with equal recommended learning opportunities, according to the platform principles, context, and pedagogy. Though the importance of ensuring equality of learning opportunities has been well investigated in traditional institutions, how this equality can be operationalized in online learning ecosystems through recommender systems is still under-explored. In this paper, we shape a blueprint of the decisions and processes to be considered in the context of equality of recommended learning opportunities, based on principles that need to be empirically-validated (no evaluation with live learners has been performed). To this end, we first provide a formalization of educational principles that model recommendations' learning properties, and a novel fairness metric that combines them to monitor the equality of recommended learning opportunities among learners. Then, we envision a scenario wherein an educational platform should be arranged in such a way that the generated recommendations meet each principle to a certain degree for all learners, constrained to their individual preferences. Under this view, we explore the learning opportunities provided by recommender systems in a course platform, uncovering systematic inequalities. To reduce this effect, we propose a novel post-processing approach that balances personalization and equality of recommended opportunities. Experiments show that our approach leads to higher equality, with a negligible loss in personalization. This paper

Work partially conducted while the author (Mirko Marras) was at University of Cagliari, Cagliari, Italy.
Work conducted while the author (Ludovico Boratto) was at EURECAT - Centre Tecnologic de Catalunya, Barcelona, Spain

This article belongs to the Topical Collection: *The FATE of AIED*
Guest Editors: Kaśka Porayska-Pomsta, Beverly Woolf, Wayne Holmes and Ken Holstein

✉ Mirko Marras
mirko.marras@acm.org

Extended author information available on the last page of the article.

provides a theoretical foundation for future studies of learners' preferences and limits concerning the equality of recommended learning opportunities.

Keywords AIED · Ethics · Learning analytics · Recommender systems

Introduction

Learning experience selection by learners is at the heart of curriculum development and, consequently, is vital to shaping individuals' knowledge and competencies (Talla, 2012; Druzhinina et al., 2018). The term *learning experience* generally refers to interactions in courses, programs, or other situations where learning takes place, including traditional and non-traditional settings (Girvan, 2018). Notable examples of the latter, with an impact on individual experiences, are online course platforms, such as Coursera and Udemy. The proliferation of initiatives and the increasing adoption of these platforms have been requiring automated mechanisms to support learning experience selection by learners, tailored to the platform's principles, context, pedagogy, and needs (Rieckmann, 2018).

One aspect receiving special attention to support the learning experience selection on these online platforms is the ranking of courses deemed of relevance to individual learners. As a result, recommender systems are being deployed to suggest courses that accommodate learner's interests and needs (Kulkarni et al., 2020). These recommended courses can be envisioned as learning opportunities being offered for the attention of a learner. Though optimizing recommendations according to learners' interests has been seen for years as the ultimate goal in the context of educational recommender systems, important principles (i.e., properties the platform aims to pursue, such as the validity, learnability, quality, and affordability of the recommended courses[1]) and the extent to which they are equally met across learners should be considered to shape online learning opportunities (Talla, 2012; Druzhinina et al., 2018). Recommendations thus need to meet a trade-off between the interests of learners and the principles of the platform, providing learners with a well-rounded range of learning experiences (Abdollahpouri et al., 2020).

Recommender capabilities represent a fundamental component of artificial intelligence systems in education. For this reason, ensuring equality among learners according to the recommended educational opportunities is essential, as the suggested courses may translate to educational gains and losses for the learners. By extension, education significantly influences individuals' life chances in the job market, and these opportunities should not be undermined by arbitrary decisions provided by a recommender system. Meyer (2016) has revealed how equal learning opportunities, equal learning outcomes, and equal job opportunities relate to each other and emphasized the indispensable, but at the same time potentially dangerous, need for equal learning opportunities. The demand for equal learning opportunities alone can lead to (i) attributing unequal learning outcomes that could have been avoided to unequal talent and effort, (ii) justify social inequalities by saying that all measures were taken to realize equal learning opportunities, (iii) limit efforts to merely

[1] "Modeling Recommended Learning Opportunity through Principles" provides a detailed description of the principles covered in this paper.

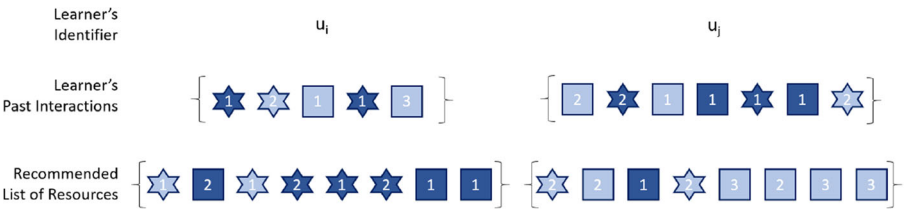


Fig. 1 Example of Inequality in Recommended Learning Opportunities. We consider two learners, u_i and u_j (first line). The mid-line shows us that the learners interacted with similar resources in terms of quality (star:high; square:low), validity (light-blue:old; dark-blue:fresh), and affordability level (1:low; 2:mid; 3:high). However, if we consider ranked lists provided by a collaborative algorithm to those two learners (bottom line), u_i's recommendation list consists of mostly fresh, high-quality, and affordable resources, while u_j's recommendations focus on obsolete, low-quality, and expensive resources

realize equal educational opportunities. These aspects have been investigated in traditional educational settings worldwide, such as in China (Golley & Kong, 2018), Germany (Buchholz et al., 2016), Japan (Fujihara & Ishida, 2016), Korea (Byun & Park, 2017), Spain (Fernández-Mellizo & Martínez-García, 2017), and United States (Shields et al., 2017).

Thanks to extensive empirical analyses, these studies have identified several variables that may lead to unequal educational opportunities, with the gap in up-to-date competencies required by the job market and the considerable costs of access to education as two of them. Operationalizing these principles and the consequent notion of equal learning opportunities in online ecosystems via recommenders is still underexplored. These systems learn patterns from data containing inherent biases, which end up being amplified in the recommendations based on such data (Boratto et al., 2019). Some learners might thus receive unequal opportunities based on the principles pursued by the targeted educational ecosystem. Figure 1 shows an example of this phenomenon[2]. Hence, it is imperative to mitigate inequalities while retaining personalization.

In this paper, we propose the concept of equality of recommended learning opportunities in personalized recommendations. To investigate how this concept applies to the online learning ecosystem, we envision a scenario wherein the educational platform should guarantee that a set of learning principles are met for all the learners, to a certain degree, when generating recommendations according to the learner's interests. We assume that those principles can be operationalized in terms of properties held by a list of recommendations. Therefore, the ideal recommender system would (i) achieve higher consistency between the principles pursued by the platform and those measured in the recommendations, (ii) retain the consistency across different learner populations while (iii) honouring individual interests. Under this scenario, we characterize the recommendations proposed to learners in a real-world online course platform as a function of seven principles derived from knowledge and curriculum literature, as reported in "Problem Formulation". Ten pre-existing available recommender algorithms, whose details are presented in "Recommendation Algorithms and Protocols", are evaluated. Our exploratory analysis sheds light on systematic

[2]Please note that the figures in this manuscript are best seen in color.

inequalities against learners based on the properties of the suggested courses. The results of our study thus motivate us to devise a novel post-processing approach that balances equality and personalization in recommendations. Specifically, the core assumption is that the notion of equality might be enhanced by trying to balance out desirable properties of course recommendations and that this objective can be achieved by re-ranking the courses originally suggested (and optimized for personalization) by the recommender system, such that the lists recommended to learners meet desirable principles of course recommendations equally across learners. The contribution of the work presented in this paper is four-fold:

1. **Operational**: we define principles that model learning opportunity properties, and we combine them in a fairness metric that monitors the equality of recommended learning opportunities across learners.
2. **Social**: we provide observations and insights on learning opportunities in recommendations, using a dataset that includes more than 40K learners and 30K courses.
3. **Technical**: we propose a post-processing recommendation approach that aims to balance personalization and equality of learning opportunities to enable optimization of different combinations of the principles of equality.
4. **Ethical**: we evaluate our approach on a real-world publicly available dataset, and we show how it may lead to higher equality of recommended learning opportunities among learners, with a negligible loss in personalization.

Our study represents a step toward understanding how equality principles can be operationalized and combined in a formal notion of equal opportunities in educational recommendations. This paper shapes a blueprint of the decisions and processes to be recommended, based on principles that need to be empirically-validated (no evaluation with live learners has been performed), and serves as a theoretical foundation for future studies of learners' concepts of fairness, preferences, or limits concerning the equality of recommended learning opportunities. For instance, this paper can be used to create examples of what questions to ask as part of interviews with learners, what scenarios to explore to elicit their concepts of fairness, and how to process data in the platform to monitor and ensure the equality principles.

The remainder of this paper is structured as follows. Section "Related Work" presents related work. Section "Problem Formulation" introduces the proposed principles and notion of equality, and "Exploratory Analysis" depicts the explorative analysis. Then, "Optimizing for Equality of Learning Opportunities" describes and evaluates our approach for mitigating inequality of recommended opportunities. Finally, "Conclusions" provides concluding remarks and discusses future research directions.

Related Work

This research lies at the intersection of Artificial Intelligence in Education (AIED), Recommender Systems (RecSys), and Fairness, Accountability, Transparency, Explainability (FATE).

Educational Recommender Systems in the Artificial Intelligence Context

The advances in the area of computing technologies have facilitated the implementation of artificial intelligence applications in educational settings, improving teaching, learning, and decision making (Pinkwart, 2016). Learners' behavioral patterns have been analyzed to make inferences, judgments, or predictions, serving for personalized guidance or feedback to students, teachers, or policymakers, for example, as proposed by Mao (2019) and Ren et al. (2019).

Our study in this paper treats recommender capabilities as a crucial component of AIED systems (Khanal et al., 2019). Given the increase in resources available in online course platforms and learning management systems, designing personalized recommendations has become a key challenge. This challenge, thereby, motivates research carried out by the AIED, Educational Data Mining (EDM), and Learning Analytics and Knowledge (LAK) communities. The most common objective in prior work is to suggest resources or peers in a given course. For instance, Lin et al. (2020) proposed a deep attention-based model to recommend resources based on learners' online behaviors. This work outperformed state-of-the-art baselines in terms of accuracy (i.e., the extent to which the recommended items are among those included in the test set for that learner, meaning that the recommender system predicts well the future interests of the learner). Similarly, Wang et al. (2019) introduced a recommendation algorithm for textbooks, showing that adding adaptivity significantly increases engagement.

Beyond recommending learning resources, Eagle et al. (2018) designed an algorithm for individualized help messages. Further, the work demonstrates that the needs of learners for a lesson can be effectively predicted from their behavior in prior lessons. Mi and Faltings (2017) and Chen and Demmans (2020) showed the important role of personalization while modeling forum discussion recommendations. Both works illustrated the presented algorithms' effectiveness in predicting learners' preferences. Additionally, Chau et al. (2018) assisted instructors with recommendations on the most relevant material to teach. Reciprocal recommendations have been investigated by Labarthe et al. (2016) and Potts et al. (2018). These works proposed two recommendation approaches for personalized contact lists. Their experiments uncovered that learners are more likely to engage in courses if they received peer recommendations. Finally, other tasks dealt with the accuracy of recommender systems in matching learners and job offers (Jacobsen & Spanakis, 2019).

Course recommendations have recently received attention due to the increasing number of initiatives carried out online. For instance, Pardos and Jiang (2020) generated course recommendations that are novel and unexpected, but still relevant to learners' interests. Their results revealed that providing services optimized for serendipity allows learners to explore resources without a strong bias towards the learner's (past) experience. Furthermore, Morsomme and Alferez (2019) found that students find recommendations for courses at other departments very helpful. In the university context, Esteban et al. (2018) and Boxuan et al. (2020) described two hybrid-methods for discovering the most relevant criteria that affect the course recommendation for university learners. Their respective results confirmed that the overall rating that a learner gives to a course is the most reliable information source. However, when it is complemented with other criteria about the courses, the

recommendation accuracy increases. Capturing the sequential relationships across courses made it possible to devise a course recommender system in Polyzou et al. (2019). Their course recommender system outperforms other collaborative filtering and baseline approaches.

Extensive research work has been devoted to mastery learning in intelligent tutoring systems, which select educational resources for learners based on knowledge tracing. For instance, Thaker et al. (2020) automatically identified the most relevant textbooks to be recommended by incorporating learner's knowledge states. Chanaa and Faddouli (2020) proposed a model that predicts learner's needs for recommendation using dynamic graph-based knowledge tracing. By learning feature information and topology representation related to learners, their model achieved a competitive accuracy of more than 80%. To avoid the mismatch between learners and learning resources, Dai et al. (2016) introduced a recommender system for suggesting learning resources, with a domain knowledge structure to connect learners' skills and learning resources. They showed that the accuracy is higher when texts related to the concerned domain knowledge are involved. In Chan et al. (2006), the authors conclude that ready-to-hand access conceives the potential for a new evolution of technology-enhanced learning (TEL) phase. This phase is defined as "seamless learning spaces" and marked by a succession of the learning experiences over different scenarios. Further, it arises from the availability of one (or more) device(s) per student ("one-to-one"). The one-to-one TEL holds the potential to "cross the chasm" from early adopters handling isolated design studies to adoption-based research and extensive implementation. Finally, Ai et al. (2019) designed an exercise recommender that considers exercise-concept mappings while tracing learners' knowledge. This recommender led to a better performance than the heuristic policy of maximizing learners' knowledge level.

Our contribution differs from prior work in three major ways. First, current approaches have been mostly optimized for learners' preference prediction, given their ratings, performance, grades, or enrolments. Conversely, our approach aims to balance how learning opportunities vary, based on high-level properties directly measurable on the ranked lists (e.g., familiarity and learnability of the recommended courses), going beyond the accuracy in predicting the future learner's preferences only. Second, even though several recommender systems integrated beyond-accuracy aspects, such as learnability and serendipity, combining them with other aspects and decoupling them from the underlying recommendation strategy appears impractical. By contrast, our post-processing mechanism can be applied to the output of any recommender system to arrange recommendations that meet a range of properties. Third, controlling how much the generated recommendations are equally consistent across learners has been rarely investigated. Hence, we introduce and operationalize a novel fairness metric that monitors equality among learners concerning the targeted educational principles.

Fairness in Artificial Intelligence for Education

Characterizing and counteracting potential pitfalls of data-driven educational interventions is receiving increasing attention from the research community. Educational

applications of artificial intelligence are not immune to the risks observed in other domains. Moreover, the design of systems may often be driven more by profit than by actual educational impact, with serious potential risks (e.g., algorithmic biases, invasion of privacy, or negative social impacts) out-weighting any benefits (Shum, 2018; Bulger, 2016; Williamson, 2017).

Responding to these concerns may be critical to determine the fairness of AIED systems and to shape how the ethics for human learning are more broadly defined. However, only a few works have focused on ethics in AIED, where the increasing use of learning analytics and artificial intelligence raises unique context-specific challenges (Ocumpaugh et al., 2014). For instance, while most existing fairness auditing and de-biasing methods require access to sensitive demographic information (e.g., age, race, gender) at an individual level, such information is often unavailable to AIED practitioners (Holstein et al., 2019a). Also, it becomes challenging to define fair outcomes in contexts where a system results in disparate outcomes across subpopulations, such as learners having lower or higher prior knowledge (Hansen & Reich, 2015). Although the community has been interested in the ethical dimensions of data-driven educational systems (Drachsler et al., 2015; Sclater & Bailey, 2015; Tsai & Gasevic, 2017), the focus has often been on policies.

Despite this widespread attention, fairness has been rarely discussed from a more practical and technical perspective (Holmes et al., 2019; Holstein & Doroudi, 2019; Mayfield et al., 2019; Porayska-Pomsta & Rajendran, 2019). Given that designing methods for addressing unfairness challenges can be highly context-dependent (Holstein et al., 2019b; Green & Hu, 2018; Selbst et al., 2019), the education research community has started to explore what fairness, accountability, transparency, and ethics look like in technology-supported education specifically. For instance, (Yu et al., 2020) found that combining the profile and material data sources does not fully neutralize biases, and it still leads to high rates of underestimation among disadvantaged groups for learners' success prediction. Similarly, (Doroudi & Brunskill, 2019) showed that knowledge tracing algorithms are susceptible to unfairness, but that knowledge tracing with the additive factor may be fairer. Hu and Rangwala (2020) focused on individually fair models for identifying students at-risk of underperforming. The work shows how to effectively mitigate bias in models and make the models useful in aiding all learners. Conversely, Abdi et al. (2020) investigated the impact of complementing educational recommender systems with transparent justifications for their recommendations. This impact leads to a positive effect on engagement and perceived effectiveness and an increasing sense of unfairness due to learners not agreeing with how their competency is modeled. Such appraisal is key to enhancing our understanding of fairness, building on knowledge gleaned from AIED research.

However, to the best of our knowledge, controlling equality in educational recommender systems has been so far under-explored. Consequently, we investigated how fairness and ethical aspects have been treated by the general-purpose RecSys community (Barocas et al., 2017; Ramos et al., 2020), analyzing whether and how the resulting treatments can be tailored to recommender systems in education. Fairness across end-users deals with ensuring that users who belong to different protected classes (group-based) or are similar at the individual level (individual-based) receive recommendations with the same quality. Group-based fairness requires that the

protected groups are treated similarly to the advantaged groups or the population as a whole. For instance, Zhu et al. (2018) designed an approach that identifies and removes from tensors all gender information about users. This approach leads to fairer recommendations (i.e., with a smaller difference in the quality of the recommendations received by user's groups) regardless of the user's group membership. Rastegarpanah et al. (2019) generated artificial data to balance group representations in the training set and minimize the difference between groups in terms of mean squared error. Similarly, (Yao & Huang, 2017) proposed metrics related to population imbalance (i.e., a class of users characterized by a sensitive attribute being the minority) and observation bias (i.e., a class of users who produced fewer ratings than their counterpart). Under a similar scenario, for instance, Beutel et al. (2019) built a pairwise regularization that penalizes the model if its ability to predict which item was clicked is better for one group than the other. These works showed that operationalizing their metrics in the recommender's objective function results in fairer recommendations.

Group-fairness may be, unfortunately, inadequate as a notion of fairness, given that there exist circumstances wherein group fairness is maintained but, from an individual point of view, the outcome is blatantly unfair. Hence, our study cares about learners as individuals, not as belonging to a class based on a certain sensitive attribute. This condition also fits with educational scenarios where sensitive demographic attributes (e.g., age, race, gender) at an individual level are unavailable to learning analytics practitioners. Examples of individual fairness notions proposed by the RecSys community (Biega et al., 2018; Lahoti et al., 2019a; Singh & Joachims, 2019) imply that similar users should have similar outcomes. Their definition of fairness states that any two individuals similar concerning a particular task should be treated likewise, assuming that a similarity metric between individuals exists. For instance, in a health-related recommender system, two patients with a similar pathology should receive recommendations of the same quality.

Our study generalizes the original definition of individual fairness and applies it to the educational context. Specifically, we aim to provide all learners, indistinctly, with recommended learning opportunities that are equally consistent with the targeted principles. We do not rely on any notion of similarity across pairs of learners based on how the targeted principles were met in the past. Compared to our definition, the other existing ones could even emphasize existing inequalities (e.g., two learners who similarly experienced less learnable recommended courses in the past could end up receiving low learnable courses more and more, though the recommender would have been fair under the original individual fairness notion). On the other hand, achieving the fairness goal indistinctly for all learners, as per our definition, can be more challenging since the demographic and behavioral (e.g., in terms of preferences) similarity between learners can vary significantly.

Problem Formulation

In this section, we formalize recommendation concepts, educational principles, and metrics that respectively monitor consistency and equality of recommended learning

opportunities among learners, according to our definition of fairness, as explained earlier.

Preliminaries

Given a set of learners U and a set of educational resources I, we assume that learners expressed their interest for a subset of resources in I. The feedback collected from learner-resource interactions can be abstracted to a set of pairs (u, i), implicitly obtained from user activity, or triplets $(u, i, rating)$ explicitly provided by learners, denoted in short by $R_{u,i}$. We denote the learner-resource feedback matrix by $R \in \mathbb{R}^{M*N}$ where $R_{u,i} > 0$ indicates that learner u interacted with resource i, and $R_{u,i} = 0$ otherwise. Furthermore, we denote the set of resources that learners $u \in U$ interacted with by $I_u = \{i \in I : R_{u,i} > 0\}$.

We assume that each resource $i \in I$ is represented by an m-dimensional feature vector $F_i = (f_1, \ldots, f_m)$ over a set of features $F = \{F_{i,1}, F_{i,2}, \ldots, F_{i,m}\}$. Each dimension F_j can be viewed as a set of values or labels describing a feature of a resource i, $f_{i,j} \in F_j$ for $j = 1, \ldots, m$. In our experiments, we considered five features, i.e., instructional level (discrete), resource category (discrete), last update timestamp (discrete datetime), number of enrolled learners (continuous), and price (continuous). Furthermore, we assume that each resource $i \in I$ is composed of a set of assets L_i. Each $l_{i,j} \in L_i$ has a type $t_{i,j} \in T$. In our study, we considered $T = \{Video, Article, Ebook, Podcast\}$, due to their popularity and their availability in the public datasets.

We assume that a recommender estimates relevance for unobserved entries in R for a given learner and uses them to rank resources. It can be abstracted as learning $\widetilde{R}_{u,i} \in [0, 1]$, which represents the predicted relevance of resource i for learner u. Given a certain learner u, resources $i \in I \setminus I_u$ are ranked by decreasing $\widetilde{R}_{u,i}$, and top-k, with $k \in \mathbb{N}$ and $k > 0$, resources are recommended. Finally, we denote the set of $k \in \mathbb{N}$ resources recommended to user u by $\tilde{I}_u$.

Modeling Recommended Learning Opportunity through Principles

Given that the recommendation capabilities are a relevant part of AIED systems, investigating whether educational recommender systems are fair and how they can be made a vehicle for making our educational systems fairer is essential. Capturing, formalizing, and operationalizing notions of equality can shape our understanding of the extent to which the educational offerings available to learners provide them with equal opportunities and how recommender systems influence the normal course of educational business. To this end, defining the variables to be equalized constitutes a natural pre-requisite.

Organizing learning opportunities in classroom settings has been traditionally a responsibility of instructional designers or teachers. To this end, they rely on a range of principles coming from the curriculum design field, including significance, self-sufficiency, validity, interest, utility, learnability, feasibility (Talla, 2012; Druzhinina et al., 2018). Hence, our study assumes that the notion of equality needs to consider these principles derived from the instructional design beliefs as those to be equalized

in recommendations, given their real-world validity for learners' educational experiences from the instructional perspective. However, we do not argue that this approach and the consequent principles are the only ones as they strongly depend on the educational context and the outcomes of the fairness auditing processes in the target context. Our principle modelling aims to serve as a starting point for researchers, to guide them in what questions, scenarios, and data to explore, while addressing the questions related to fairness. Therefore, we argue that our study offers a blueprint for the decisions and processes needed.

Human inspection of curriculum-design-based principles is usually based on textual guidelines, and the translation into numerical indicators, when available, is dependent on the specificity of the educational context. Given the unique characteristics of the online educational context and the constraints the platforms introduced in the collection of learners' data, we assume that the principles are based on data that would typically be available in a platform in which an educational recommender system would be embedded. For this reason, not all the principles and not all the guidelines can be directly operationalized[3]. One of the core assumptions of our approach is then that it should be possible to define those principles in terms of properties held by a list of recommendations. Specifically, we envision a scenario wherein only a representative subset of principles are embedded in the recommender system's logic. The educational platform is thus empowered with the capability of controlling the extent to which the list of courses recommended to learners meets each principle. While the high-level conception of the selected principles is assumed to be relevant, their operationalization into the recommender's logic is dependent on the platform, turning to simplified implementations in some cases. While we provide formulations that are as general as possible, we will ensure that our approach can be extended or adapted to any principle.

Formally, we consider a set C of functions $c_{\tilde{I}_u}(\cdot) : I^k \rightarrow [0, 1]$. Each function receives a set of k resources I^k and returns a value indicating how much the set of resources meets that principle. The higher the value, the higher the extent to which the principle is met. Specifically, we consider the following seven principles, whose mathematical formulation is provided in Appendix A.

Definition 1 (Familiarity) *Familiarity is defined as whether the learner is familiar with the recommended content, as measured by whether the relative frequency of the course categories in a recommended set is proportional to that in the courses the learner took.*

Familiarity is at the heart of learner-centered education. Learners might be more comfortable if the subject matter is meaningful to them, and it is assumed that it becomes meaningful if they are familiar with that subject. Xie and Joo (2009) supported this observation through descriptive and statistical analysis, uncovering that the familiarity was correlated to the content searching behavior. Similarly, Qiu and Lo (2017) showed that participants were behaviourally and cognitively more engaged

[3] "Limitations" identifies a range of restraints derived from these assumptions.

in tasks with familiar topics as well as having a more positive affective response to them.

Our study models familiarity using the category of the resources in a recommended list, encoded into a pre-defined taxonomy. If the relative frequency of the course categories in a recommended set is proportional to that in the courses the learner took, we assume that the familiarity is high (a value of 1). Conversely, the minimum familiarity of 0 is achieved when the recommender suggests resources in the opposite direction concerning the learner's most familiar categories. This principle is related to the concept of calibrated recommendations, which aim to reflect the various interests of a user in the recommended list with their appropriate proportions (Steck, 2018)[4].

Definition 2 (Validity) *Validity is defined as whether the course is likely to be up-to-date and not obsolete, as measured by when content was last updated. A subject is assumed to be more valid if it has been newly updated.*

Controlling the validity of the learning content is one of the major axes of education since learners would not find information invalid anymore in the courses. One way of maintaining the validity of the course content is to continuously update it, either with more recent content or with new versions of the same contents (e.g., adapted based on the learners' feedback). This practice also shows learners that the course is alive. Curriculum-design experts usually seek to follow current trends and carefully consider the validity of a curriculum (Druzhinina et al., 2018); otherwise, the opportunity becomes obsolete. Similarly, Bulathwela et al. (2019) highlighted that content freshness is one of the main factors shaping content validity. Hence, we assume that validity should be to be taken into account in the recommendations offered.

Our scenario operationalizes the validity principle by controlling that learners are presented with recommended courses that have been recently or frequently updated. Values close to 0 imply that the recommended list includes courses no longer updated for a long time, while values close to 1 are achieved by recommended courses with recent updates[5].

Definition 3 (Learnability) *Learnability is defined as whether the recommended courses present an opportunity that is coherent with the learners' ability, with the learnability measured as whether the set of courses varies in terms of instructional level.*

Learnability is generally associated with the ease, efficiency, and effectiveness with which learners can perform a knowledge acquisition activity. Our study assumes

[4]Differently from that work, which used the Kullback–Leibler divergence as a non-symmetric, unbounded, and computationally unstable distance function, we adopted the *Hellinger* distance, which is symmetric and bounded in the range [0, 1].

[5]It should be noted that using recency of updates as a proxy for validity does not consider that, for instance, a course on foundational material updated many times in the past does not benefit from recent updates.

that the subject matter to be recommended should be within the knowledge schema of the learners. The literature indicated that learnability impacts learner motivation to learn (Conaway & Zorn-Arnold, 2016), prompting us to monitor this principle in the recommended lists.

In our scenario, this concept is operationalized to ensure that courses of diverse instructional levels are presented and maximize the possibility that learners can find an opportunity coherent with their abilities. Please, note that our study is not learner-centric, i.e., no record of student learning, performance, or exam grades is made and, therefore, student knowledge is not tracked. The factors tracked to measure learnability are the instructional levels of the courses attended by and recommended to learners. Compared to knowledge-tracking methods, our operationalization might appear an over-simple way of modelling the zone of proximal development, i.e., the zone between the actual level of development of the learner and the next level attainable through the use of mediating tools and/or collaboration, defined by Vygotsky (1978). However, the current online course platforms impose constraints that should be met. Specifically, data on mid-term quizzes and final exams are often not recorded internally, and this leaves the implementation of traditional knowledge-tracking techniques in these platforms as an open challenge. Hence, we rely on course recommendations that cover different instructional levels. Learnability values close to 0 imply inequality among levels, while the high balance is obtained with values close to 1.

Definition 4 (Variety) *Variety is defined as whether the recommendation takes into account that learners are different and learn in different ways based on their interests and ability, as measured by the degree to which the recommended courses present a mix of different asset types.*

Providing course material in a variety of formats represents a primary objective. For instance, by studying the online course design and teaching practices of award-winning teachers, Kumar et al. (2019) uncovered that including video, audio, reading, and interactive content made courses more engaging and appreciated by learners during their learning sessions, though no explicit mention of the effectiveness of variety on learning gains has been made. In another study, Papathoma et al. (2020) highlighted how this variety of formats increases the accessibility of a course, given that learners may struggle with a particular medium (e.g., due to a reading barrier such as dyslexia or a video barrier such as hearing or attention problem). Therefore, monitoring whether a course provides learners with a large variety of content formats is crucial.

Our operationalization of variety assumes that varied asset types may be provided to help learners comprehend the subject from various perspectives. Hence, variety values close to 0 mean that the learning opportunities are focused on only one asset type, while types greatly vary for values close to 1.

Definition 5 (Quality) *Quality is defined as the perceived appreciation of the recommended resources by the learners, as measured by the ratings that the learners assign to resources after interacting with them.*

Student evaluation of teaching quality is prominent to assess current teaching experiences. Teaching evaluation helps to promote a better learning experience for learners and provide information to future learners while deciding for attending a course. However, defining quality in online learning is challenging because there is no real consensus on its true meaning. Consequently, quality is evaluated differently depending on the organization in charge of measuring it. For instance, Darwin (2017) showed that student ratings are perceived as a valuable, though fragile, source of intelligence about the effectiveness of curriculum design, teaching practices, and assessment strategies. On the other hand, Gómez-Rey et al. (2016) observed that learners considered other core variables in defining quality in online programs, such as the ability to transfer, knowledge acquisition, learner satisfaction, and course design.

Our study operationalizes quality by leveraging the learners' ratings. Rating values close to 0 mean that the learning opportunities are of low quality (i.e., they have received a low rating from other learners), while values close to 1 are measured for high-quality recommended opportunities. Though some studies demonstrated that learners' ratings do not often correlate with other measures of quality (e.g., learning outcomes), this design choice makes it possible to meet the current constraints in data gathering in large-scale educational platforms and allows us also to maintain this principle as general as possible.

Definition 6 (Manageability) *Manageability is defined as whether the online classes are large or small, as measured by the number of learners enrolled in the recommended courses, with small classes considered more manageable.*

Organizational aspects are critical for shaping learners' experiences. In this context, class size differences may influence academic interactions between students and their professors and peers. For instance, with a large number of learners, the instructor may work harder to combat student passivity and encourage participation, as learners feel an increasing sense of anonymity. This point is confirmed by the study of Beattie and Thiele (2016), which uncovered that the likelihood of academic interactions about course material and assignments with professors was diminished in larger classes, as was the probability of talking to peers about ideas from classes. Similarly, Lowenthal et al. (2019) revealed that online courses with fewer enrollments are seen better for student learning and faculty satisfaction by learners and instructors.

Our study embeds the notion of manageability, associating it with the size of the course class where the recommended opportunities take place. This principle is relevant to offer opportunities under smaller and controlled classes. Hence, manageability values close to 0 mean that the learning opportunities include very large classes, while values close to 1 refer to small classes[6].

[6]It should be noted that other operationalizations might refer to a teacher's ability to manage students or how topics are selected by teachers. Other aspects of manageability (e.g., number of teaching assistants, number of recitation sections) can be captured under this principle, depending on the educational context and platform capabilities.

Definition 7 (Affordability) *Affordability is defined as the cost of accessing the recommended opportunities, as measured by the enrolment fees of the suggested courses, with less expensive courses having higher affordability value.*

Dealing with the increasing costs of education is critical, given that lots of learners need access to vastly more affordable and quality education opportunities, including tuition-free course options. For instance, Mohapatra and Mohanty (2017) emphasized that the affordability of the offering is one of the prime predictors of the learners' perception, while Joyner et al. (2016) uncovered that providing more affordable courses has led to the learner population that is more intrinsically motivated to learn, more experienced, and more professionally diverse in some contexts. Institutions and platforms are thus under increasing pressure to provide more affordable learning without sacrificing optimal learning outcomes. For this reason, we monitor the affordability principle in the recommendations generated in the online platform.

Our notion of affordability aims to control the degree of economic accessibility for the recommended opportunities, measured by their enrolment fees. Specifically, we consider how much the learning opportunities cover a range of fees. A value close to 0 means that the learning opportunities are expensive, while a value close to 1 corresponds to free-of-charge opportunities.

Though each of the principles has relevance for students' educational experiences from the instructional design perspective, the set of principles could be expanded. Additionally, the proposed set is not meant to be the unique right set. Furthermore, it should be noted that massive online course platforms are often targeted for profitability and large coverage, and a business plan should be provided and making a profit must be considered as one of the primary goals. Thus, the question of how to integrate business and educational principles remains an open one. This question deserves a broader and specific discussion, going beyond a closer inspection of the technicalities. Our study in this paper assumes that the educational platform aims to impact the learners positively. Furthermore, there might be several principles left out, but relevant for certain educational scenarios or specific platforms (e.g., the time of day a course is offered). This observation challenges an assumption that there is a one-size-fits-all set of principles to be equalized in educational recommender systems. Another point to mention is that the considered principles assume a top-down approach and seem to leave little in the way of learner autonomy to help in their decision-making about the courses they take. However, this is not fully the case, given that recommendations are meant as a suggestion to learners, and learners are the entity that makes the final decision on the courses to attend. In addition, the principles to be considered and the importance to give to each of them could be tailored to each learner individually, through ad-hoc user interfaces integrated into the course platform. The protocols and interfaces required to favor customization at the individual learner level go beyond the scope of this paper, although our approach might be adapted.

Equality of Recommended Learning Opportunity

To formalize the equality of recommended learning opportunities, we first need to define how much the list recommended to each learner meets the principles targeted

by the educational platform. In this paper, we propose to operationalize the concept of consistency across principles as the similarity between (i) the degree to which all principles are met into the recommended list and (ii) the degree of importance for the principles targeted by the educational platform. The higher the similarity, the higher the extent to which the principles are met. We resorted to the operationalization of this metric locally on each ranked list so that it will be possible to optimize such a metric on a pre-computed recommended list through a post-processing function (see "The Post-Processing Approach Proposed"). For the ranked list of recommended courses $\tilde{I}_u$ to a learner u, we assume the platform aims to ensure a targeted degree $p_u(m)$ for each principle $m \in C$ for each learner. In other words, $p_u(m)$ defines the extent to which the platform seeks to meet that principle m. The higher the $p_u(m)$ score is, the more important principle m is for the platform. The main motivation behind this term is that principles might have different importance, and the term we are defining here allows us to model the degree to which each principle should be met.

$$p_u(m) \in [0, 1], \quad \forall m \in \{0, \cdots, |C| - 1\} \tag{1}$$

Once a recommender computes the top-k resources $\tilde{I}_u$ to be suggested to learner $u \in U$, we need to define the extent to which each principle $m \in C$ is met in $\tilde{I}_u$. To this end, we measure the degree the principle m is met in the recommended list $\tilde{I}_u$ as $q_{\tilde{I}_u}(m)$. The way this score is obtained depends on the principle under consideration and how it has been operationalized (see Section 3.2 for a textual description of each principle m and Appendix A for the mathematical formulation of each principle m to obtain c). For instance, given a recommended list $\tilde{I}_u$ and assuming that m is the principle of affordability, the score c represents the extent to which the courses in $\tilde{I}_u$ are affordable (the more affordable the recommended courses are, the higher the c score is). This score is needed to monitor the gap between the degree $q_u(m)$ the principle m is met in the recommended courses and the targeted degree $p_u(m)$ expected by the platform (defined in (1)). The score $q_u(m)$ is defined as follows:

$$q_{\tilde{I}_u}(m) = c_{\tilde{I}_u}(m) \in [0, 1], \quad \forall m \in \{0, \cdots, |C| - 1\} \tag{2}$$

where the value corresponding to each principle $q_{\tilde{I}_u}(m)$ is computed by applying the formulas formalized in Appendix A.

Once we have formulated the degree $q_u(m)$ the principle m is met in the recommended courses and the targeted degree $p_u(m)$ expected by the platform, we need to define how to measure the gap between these two degrees across principles. This is of fundamental importance to assess how far the recommender system is from achieving the targeted degree pursued by the platform. Specifically, for the ranked list of a learner u, the principles targeted by the educational platform are met if the values in p_u (targeted degree of the platform) and $q_{\tilde{I}_u}$ (degree achieved in the recommended list) are aligned with each other. To assess the extent to which the principles' goals targeted by the educational platform are met (are consistent between each other), we compare the vectors p_u and $q_{\tilde{I}_u}$, measuring the distance between the two. We define the notion of **Consistency** between (i) target principles and (ii) the extent to which the principles are achieved in recommendations, by the complement of the

Manhattan (M) distance[7], a symmetric and bounded distance measure. The higher the distance is, the lower the consistency score for the target principles is. Computing this consistency score $Consistency(u|w)$ for all learners allows us to compare the extent to which recommendations are equally consistent across all learners, i.e., whether the notion of equality of recommended learning opportunities defined in our paper is met. The consistency for each learner and the entire learners' population is formulated as follows:

$$Consistency(u|w) = 1 - M(p_u, q_{\tilde{I}_u}|w) = 1 - \frac{1}{|\tilde{I}|}\sum_{i=1}^{|\tilde{I}|} w_i \left|[p_u]_i - [q_{\tilde{I}_u}]_i\right| \quad (3)$$

$$Consistency(U|w) = \frac{1}{|U|}\sum_{u\in U} Consistency(p_u, q_{\tilde{I}_u}|w) \quad (4)$$

where w is a vector of size $|C|$; the element w_i is the weight assigned to the principle i, between 0 and 1. *Consistency* is 1 if p_u (targeted degree of the platform) and $q_{\tilde{I}_u}$ (degree achieved in the recommended list) are perfectly balanced, meaning that the principles pursued by educational platform are met. Conversely, the lowest *Consistency* 0 is achieved when p_u assigns value 0 to every principle that $q_{\tilde{I}_u}$ assigns value 1 (or vice versa), so that the distributions are completely unbalanced. In the latter situation, the recommender suggests resources opposite to the educational platform's goals. Given that principles are context-sensitive, our notion of consistency might provide different target degrees of principles for each learner or each time period. For instance, concerning familiarity, different learners might have a different propensity for familiar content, and the same learner may, at different times, have distinct preferences. The above formulation allows modeling these circumstances.

Given the notion of consistency, we can formalize the notion of **Equality** across consistencies as the complement of the Gini index[8] over the consistencies across learners. The Gini index ranges between 0 and 1, with higher values representing distributions with high inequality. It is used as:

$$Equality(U|w) = 1 - GINI\left(\{Consistency(u|w) \mid \forall u \in U\}\right) \quad (5)$$

where a value of 0 represents the largest inequality across consistencies (i.e., the extent to which the degree of the principles targeted by the platform and the degree of the principles achieved by the recommender system are the same), and a value of 1 means that the recommender systems are perfectly equal across learners in terms of consistency. Differently from Lahoti et al. (2019b) and Biega et al. (2018), we count as a positive effect when learners achieve high consistency in recommendations,

[7]The Manhattan distance between two vectors is equal to the one-norm of the distance between the vectors. Specifically, it represents the distance between two points in a grid-based on a strictly horizontal and/or vertical path, i.e., along the grid lines.

[8]The Gini index is a statistical measure of distribution that aims to model inequality among a population. The coefficient ranges from 0 to 1, with 0 representing perfect equality and 1 representing perfect inequality. Formally, it is defined based on the Lorenz curve, which plots the proportion of the total consistency of the population cumulatively earned by the bottom x% of the population. The Gini index is the ratio of the area that lies between the line of equality and the Lorenz curve over the total area under the line of equality.

regardless of the consistency in their past interactions. Thus, the ideal recommender system would be the one that (i) achieves the highest consistency between the principles pursued by the platform and those measured in the recommendations, (ii) keeps it equal over the learners' population, and (iii) retains individual interests of learners.

It should be noted that our notion of equality is defined as providing the same consistency on principles to all learners, without leveraging any information on learners' sensitive features, e.g., gender. The targeted degree for each principle for each student is assumed to be set by or known by the educational platform. Our approach enables a platform to set the same targeted degree for all students or apply student-specific targeted degrees set based on the previous learners' preference or elicited from learners. To focus better on the core contribution of this paper, "Exploratory Analysis" will investigate whether all the principles can be maximized for all learners, leaving student-specific targeted degrees as part of a human-centered study[9]. Furthermore, the reliance on stakeholders empowered with decision-making capabilities to configure the platform with the considered principles and their different targeted degrees represents an essential element towards implementing our notion of equality. This primary responsibility of stakeholders is in addition to all the others involved in the educational ecosystem (e.g., selecting the preferred system, deciding the recommendation strategy, and defining the visual interface).

Exploratory Analysis

To illustrate the trade-off between learners' interests and the considered principles and further emphasize the value of our analytical modeling, we characterize the learning opportunities proposed by ten algorithms to learners of a real-world educational dataset as a function of the proposed principles.

Data

We analyze data from the educational context, exploring the role of the proposed principles in recommendations. We remark that the experimentation is challenging because there are very few large-scale educational datasets coming from this specific field of online education. To the best of our knowledge, COCO (Dessì et al., 2018) is the widest educational dataset with all the attributes required to model the proposed principles and with enough data to assess performance significantly. Collected from an online course platform, it includes 43,045 courses and 4,123,127 learners who gave 6,564,870 ratings. Other educational datasets proposed by Feng et al. (2019), Zhang et al. (2019), and Qiu et al. (2016) generally include (*learner*, *course*, *rating*) triplets only, as needed in traditional recommendation scenarios.

[9]We argue that quantifying student-specific targeted degrees from the previous learners' preferences encoded in the training set might be unreliable, given that the preference of each learner might have been biased by the recommender system itself.

Recommendation Algorithms and Protocols

We considered ten methods and investigated the recommendations they generated. Two of them are baseline recommenders, and the other eight are state-of-the-art algorithms, chosen due to their performance, wide adoption, and core applicability in learning contexts (Kulkarni et al., 2020). These algorithms are:

- Non-Personalized: Random and TopPopular.
- Neighbor-based: UserKNN and ItemKNN (Sarwar et al., 2001).
- Matrix-Factorized: GMF (He et al., 2017), NeuMF (He et al., 2017).
- Graph: P3-Alpha (Cooper et al., 2014) and RP3-Beta (Paudel et al., 2016).
- Content: ItemKNN-CB (Lops et al., 2011).
- Hybrid: CoupledCF (Zhang et al., 2018).

Based on hyperparameter tuning, UserKNN and ItemKNN relied on the cosine metric and 100 neighbors. GMF and NeuMF used 10 factors and were trained on 4 negative samples per positive instance. This means that, for each observed user-item interaction, we added to the training set four user-item pairs where the selected item has been never observed by that user in the dataset. P3-Alpha was executed with 0.8 alpha and 200 neighbors, while RP3-Beta adopted 0.6 alpha, 0.3 beta, and 200 neighbors. ItemKNN-CB mapped course descriptions to Term-Frequency Inverse-Document Frequency (TF-IDF) features. The TF-IDF features of courses into the user's profile were averaged, and their cosine similarity with the TF-IDF features of other courses is used during ranking. CoupledCF embedded user-item associations, the user tendency to interact with each category of courses, and the category of the course in the current user-item pair. To be as close as possible to a real scenario, we used a fixed-timestamp split (Campos et al., 2014). The basic idea is to choose a single timestamp that represents the moment in which test learners are on the platform waiting for recommendations. Their past corresponds to the training set, and the performance is evaluated with data coming from their future. In this work, we select the splitting timestamp *2017-06-08*, which maximizes the number of learners involved in the evaluation, by setting two constraints: the training set must keep at least 4 ratings per user, and the test set must contain at least 1 rating. This split led to 43,021 learners, 24,321 courses, and 529,857 interactions (Fig. 2). Normalized Discounted Cumulative Gain (NDCG)[10] is used as an effectiveness metric. As a measure of relevance for NDCG, the binarized (u, i) tuples formalized in "Problem Formulation" were used[11].

[10]Discounted cumulative gain (DCG) is a measure of ranking quality. The premise of DCG is that highly relevant courses appearing lower in a recommend list should be penalized, given that the graded relevance is logarithmically proportional to the ranking position. To let DCG be independent of the ranking length, DCG is normalized by scaling the results based on the best possible value, i.e., ideal DCG. The latter is computed by sorting all relevant courses in the test set by their relevance, producing the maximum possible DCG.

[11]It may happen that many more courses would be relevant to a given learner than that learner will have interacted with. Some courses that end up high on a recommended list for a given learner but that this learner did not see or have time for it would not be relevant. Dealing with this well-known problem of missing-not-at-random interactions is an open problem in the recommendation landscape (Nakagawa & Freckleton, 2008).

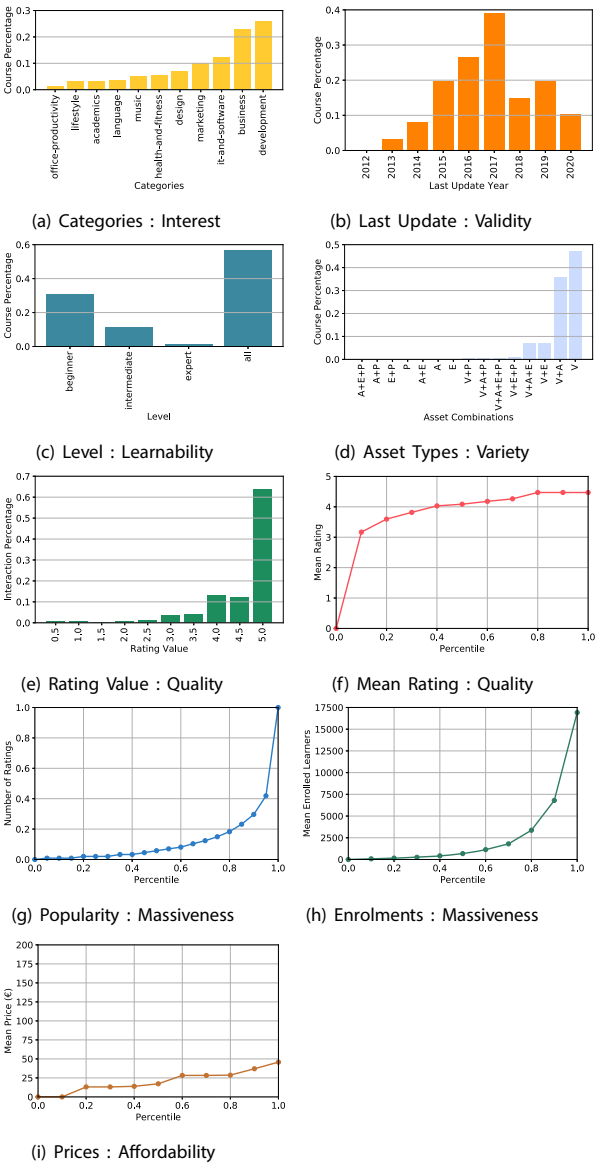


Fig. 2 Data Statistics. Characteristics of the real-world dataset relevant to the learning opportunity principles proposed by this paper: course popularity, rating values, last update timestamp, thematic category, instructional level, asset types (V:Video; A:Article; E:Ebook; P:Podcast), prices, number of enrolments per course, and average rating per course. Subfigure captions specify the feature and the interested principle as $< Feature >:< principle >$

Real-World Observations

We characterize how the proposed principles were met in the lists of courses suggested by the algorithms considered. Student-specific targeted weights for each

principle would be elicited through user groups, surveys, or implicit preferences observed in the collected data. However, due to the absence of this form of feedback in COCO and given that the preference of each learner derived from historical data might have been biased by the recommender system itself, we consider a scenario where the educational platform aims to maximize all the targeted principles, i.e., $p_u = \mathbb{1}^{|C|}, \forall u \in U$. To this end, we assume to give the same maximum weight to all the principles, i.e., $w_i = \mathbb{1}^{|C|}, \forall u \in U$. While this assumption comes with some limitations described in "Limitations", given that each learner does not always prefer maximum familiarity, for example, such setup allows us to quantify the extent to which each principle is met. We leave experiments on learner-specific weights elicited through interviews or surveys as future work. Three research questions drove our analysis:

RQ4.3.1 Does a relation exist between consistency and equality?
RQ4.3.2 To what extent principles impact on consistency and equality?
RQ4.3.3 Are consistency and equality affected by the past learners' behavior?

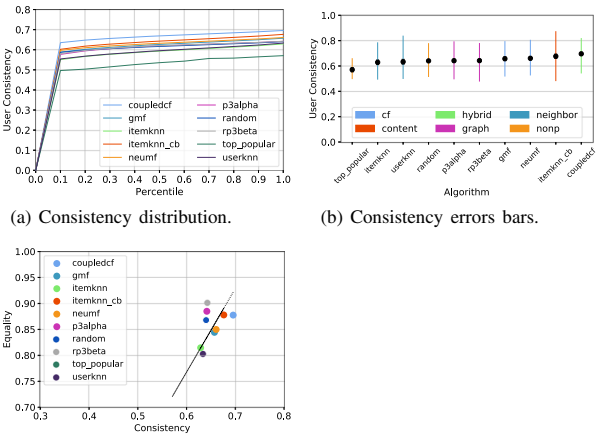
Equality Analysis (RQ4.3.1) In this subsection, to answer the first research question, we explore whether a relation between consistency and equality exists and, if this is the case, which type of relation exists. Answering this question can allow us to uncover a link between a metric that requires knowledge about the whole learner population (i.e., equality) and a metric that can be directly optimized on a single ranked list (consistency), making it possible to apply a non-NP-Hard re-ranking procedure to increase equality in our task. To this end, we provided recommendations to all learners, suggesting to each learner $k = 10$ courses; then we measured consistency across the whole learners' population, i.e., how much the principles were met in the recommendations of learners (4), and equality, i.e., how similar were the consistencies across learners (5). Furthermore, to assess the extent to which the recommender system is accurate (i.e., predicts well the future interests of the learners), we also computed the accuracy of the recommender system in terms of Normalized Discounted Cumulative Gain (NDCG). Table 1 reports the consistency, equality, and accuracy of the ten recommender systems considered. A higher value indicates that a recommender better drives consistency, equality, and/or accuracy respectively. A first observation from Table 1 is the following:

> **Observation 1**. *Recommenders that embed content metadata ensure higher equality across learners. When the recommender uses only user-item interactions, the equality is reduced. This holds regardless of the algorithm's subfamily.*

Though the observation above holds under our setting, the values associated with the equality of the recommender systems and the mean consistency values associated with each principle do not reveal much about how consistency estimates are equal across individual learners. Therefore, we plot consistencies across learners for each

Table 1 Global Indicators. Normalized Discounted Cumulative Gain (NDCG), consistency across the whole learners' population , and equality produced by different families of recommenders. Italic values highlight the highest value for each metric across algorithms. The highest NDCG is achieved by ItemKNNCB, while the highest consistency and equality was observed for CoupledCF

	Non-Personalized		Neighborhood		Graph		Matrix-Factor		Content	Hybrid
	Random	TopPopular	UserKNN	ItemKNN	P3Alpha	RP3Beta	NeuMF	GMF	ItemKNNCB	CoupledCF
NDCG	0.000	0.035	0.012	0.021	0.001	0.000	0.008	0.010	*0.042*	0.013
Consistency	0.586	0.516	0.618	0.615	0.578	0.572	0.662	0.652	0.699	*0.717*
Equality	0.872	0.795	0.885	0.891	0.850	0.847	0.917	0.906	0.959	*0.969*



(a) Consistency distribution.

(b) Consistency errors bars.

(c) Equality w.r.t. Consistency.

Fig. 3 Consistency over the Entire Population. On the left plot, lines represent the consistency distribution over learners, sorted in increasing order. On the center plot, each error bar includes mean (dot), std deviation (black solid line), and min-max values (colored thick line). The right plot highlights the direct relation between consistency and equality

algorithm, sorted by increasing values (Fig. 3a). It can be observed that ItemKNN-CB and CoupledCF are equally consistent across learners. This result might depend on the fact that, in the presence of principles related to the course content, the content-based and hybrid methods may, incidentally, increase those principles and lead to higher consistency. In other words, their equality could be biased by the fact that they capitalize on input information that is related to some principles.

While it may happen that certain principles are optimized by a traditional recommendation algorithm involuntarily, it is generally impractical to arrange the internal logic of an algorithm a priori to all the principles targeted. Figure 3b plots the consistency error bars for each algorithm, with mean, standard deviation, minimum, and maximum values. We observe that there is a link between the magnitude of the mean and the standard deviation. More precisely, the higher the mean consistency guaranteed by the algorithm, the lower the standard deviation across consistency values is (Fig. 3c). Hence, we can draw a subsequent observation:

Observation 2. *Recommenders with high consistency lead to higher equality of recommended learning opportunities. This property is stronger for neural collaborative, content-based, and hybrid recommenders.*

Uncovering a link between a metric that requires knowledge about the whole learner population (i.e., Equality in (5)) and a metric that can be directly optimized on a single ranked list (i.e., Consistency in (4)) makes it possible to apply a non-NP-Hard re-ranking procedure to solve our task. This suggests that we should investigate the interplay between (i) the average consistency across principles and (ii) the consistency achieved for each principle individually when a given learner and algorithm are considered.

We can conclude that a relation between consistency and equality exists in all the recommender systems considered. The relation is direct, i.e., the higher the consistency is, the higher the equality is, meaning that recommender systems that achieve higher consistency also tend to equalize it across learners. The strength of this relation depends on the recommender system.

Individual Principle Analysis (RQ4.3.2) Next, to answer the second research question, we investigate the extent to which the considered principles impact on consistency and equality and whether this impact is different based on the principle. An exploration of this perspective can inform us on the extent to which each principle is met and, by extension, provide helpful insights for the approach we will develop to increase equality. For the sake of readability and conciseness, we do not further consider the Random algorithm over the analysis. Figure 4 reports the mean, standard deviation, minimum, and maximum values over each principle on that recommender. For instance, the `coupledcf` plot shows that the `familiarity` score has a mean of 0.80, a standard deviation of ± 0.05, and spans the whole range (min: 0.00; max: 1.00). The first observations can be made for the top popular algorithm, whose results reveal that popular courses are mostly fresh (high validity) and have high quality. However, the consistency of these two principles comes at the price of low familiarity, learnability, variety, and affordability. Considering algorithms that capitalize on course metadata (CoupledCF and ItemKNN-CB), similar patterns can be observed across principles, except on variety and quality. For the variety and quality, embedding user-item interactions in CoupledCF can reduce the min-max gap. Hence, we can avoid situations where few learners have very high/low values. Other algorithms achieved a more stable consistency.

To assess whether certain algorithms favor or hurt a given principle, Fig. 5 reports for each principle how it varies over algorithms. We observe that familiarity and affordability suffer from high deviations, while more stable values were measured for other principles over algorithms. We conjecture that the stability observed on quality comes from the highly unbalanced rating value distribution. Indirectly, this effect could come from the fact that learners tend to evaluate courses with high ratings when they decide to rate them. Figure 5 also confirmed this intuition. On principles like affordability, manageability, and learnability, the considered algorithms got lower values.

> **Observation 3**. *Quality, validity, and manageability are guaranteed to a high extent by different recommenders, regardless of the family. Familiarity, affordability, learnability, and variety experience low absolute values and substantial deviations over algorithms, independently of the algorithm's subfamily.*

To further confirm the role of each principle over consistency, we looked at the correlation between the consistency achieved for a given principle and the consistency achieved by including all the principles. In Fig. 6, we report the results for each principle and algorithm pair. Values higher than 0 are expected when the consistency at the principle level is directly related to the high consistency achieved when all the

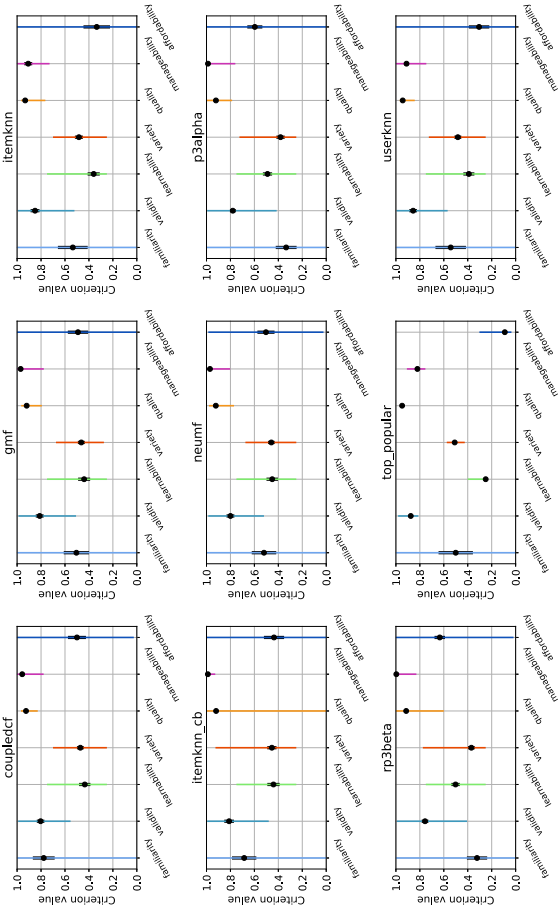


Fig. 4 Algorithm over Principle. For each recommendation algorithm, the corresponding plot reports an error bar for each principle as measured for that algorithm, including mean (dot), std deviation (solid black line), and min-max values (thick colored line)

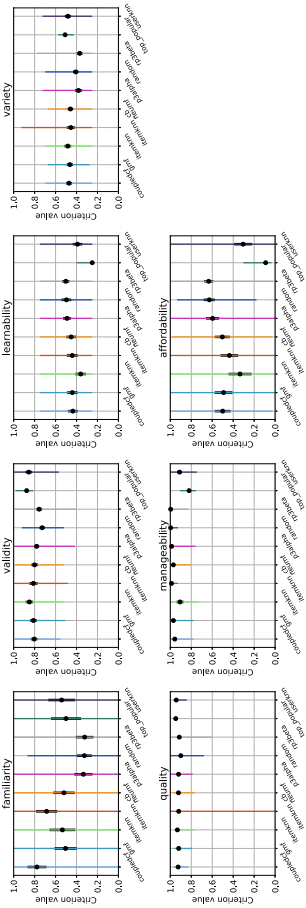


Fig. 5 Principle over Algorithms. For each principle, the corresponding plot reports an error bar for each algorithm as measured for that algorithm, including mean (dot), std deviation (solid black line), and min-max values (thick colored line)

principles are considered. Hence, the overall consistency is more likely to be met when that specific principle is met. Values lower than 0 result in the opposite behavior. No relation is found when the value is close to 0. This property allows us to make another observation:

> **Observation 4**. *Familiarity, learnability, and affordability are the most influencing principles on the overall consistency across principles. This effect is stronger for content-based and hybrid recommenders.*

We can conclude that the extent to which the principles impact consistency and equality depends on the recommender system and the principle considered. In general, certain principles (e.g., familiarity, learnability, and affordability) appear as the principles with the highest impact on consistency and equality, across all recommender systems. Those principles might be the ones that will be impacted the most by an approach that increases equality.

Past Interaction Analysis (RQ4.3.3) The last research question in this section is related to an exploration of the extent to which consistency and equality are affected by the past learners' behavior. Most of the observations seen so far are based on the fact that the observed consistency values are averaged over learners. However, it is interesting to ask whether, for two learners with similar past interactions concerning the considered principles, we should expect a similar consistency. In other words, we

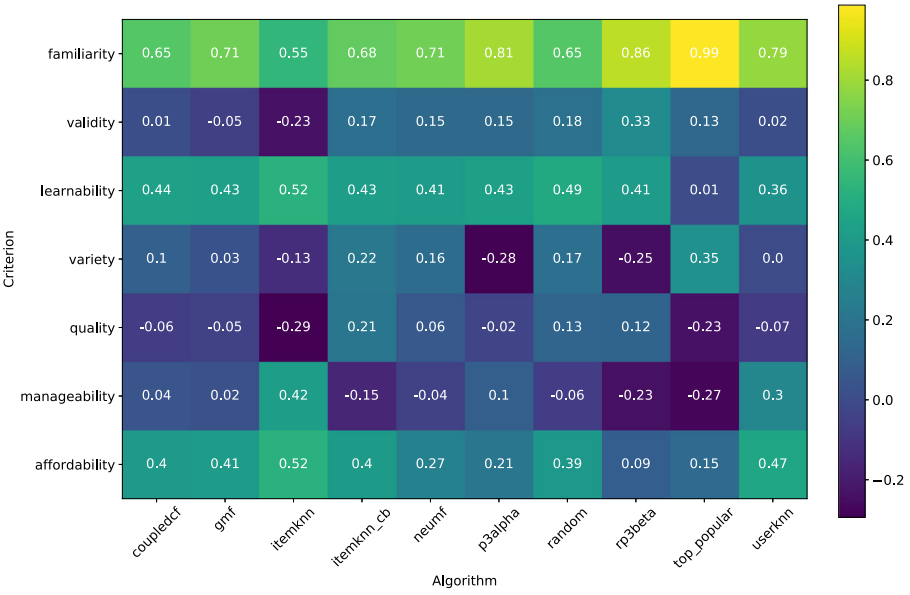


Fig. 6 Principle-Consistency Relation. Heatmap of correlations between the consistency for a given principle and the consistency for the whole principles list, over different algorithms. Each value ranges in [-1, +1], and for each principle and algorithm, the Spearman correlation is computed over a distribution of (principle value, user consistency) pairs

ask whether similar learners get similar consistency. In our setting, for learners, we assume that being similar means having similar consistency in their past interactions. Therefore, we computed the consistency metric defined in (3) by substituting the vector $q_{\tilde{I}_u}$ (i.e., the extent to which principles are met in the recommendations $\tilde{I}_u$) with the vector q_{I_u} (i.e., the extent to which principles are met in the list of courses I_u previously attended by the learner), so that we can quantify how much the targeted principles were met in the set of past interactions of each learner.

To this end, for all the possible pairs of learners, u_1 and u_2, we computed the difference of consistency in their profile and their recommendations. Figure 7 depicts pairs of results by increasing the difference of consistency in their profiles. It can be observed that, except for the graph-based P3Alpha and RP3Alpha, a higher similarity of consistency between the profiles results in a higher similarity of consistency over the recommendations. Figure 8 also shows, for each principle and algorithm, the best and worst consistency across learners, according to the above definition. It is confirmed that familiarity, learnability, variety, and affordability play a key role in overall consistency.

Observation 5. *Learners who interacted with courses aligned with the principles are likely to receive recommendations that meet those principles. Similar learners in terms of consistency in the courses they took are likely to receive a similar treatment in terms of future consistency.*

We can conclude that consistency and equality are affected by the prior courses the learner has attended in the platform. Learners whose former courses already achieve

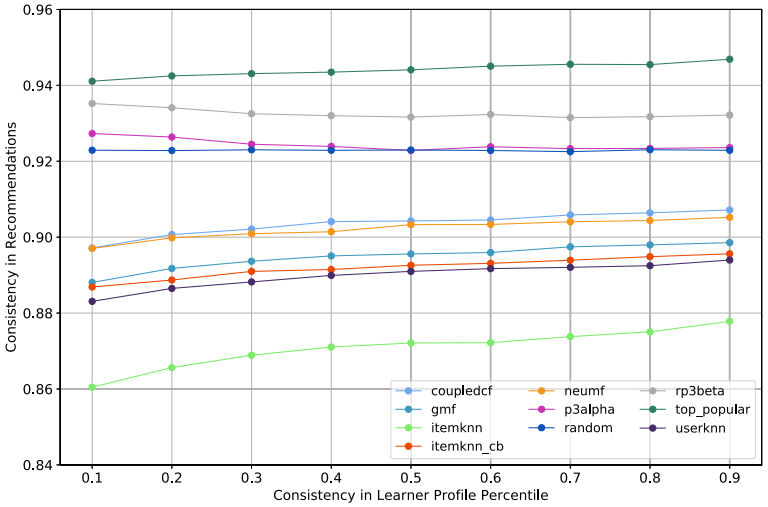


Fig. 7 Consistency in Profile and Recommendation. The lines show the difference in recommendation consistency over random pairs of learners, with values sorted by increasing difference in consistency in profiles

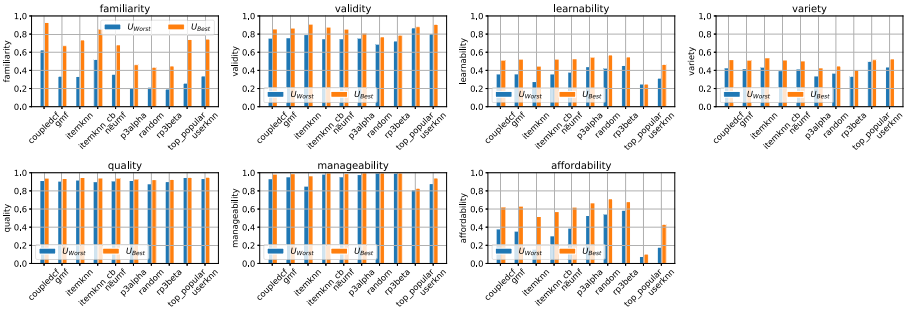


Fig. 8 Learners with More (Less) Consistency. For each principle, the corresponding plot reports the mean consistency achieved in recommendations by learners with high consistency in their profile (orange) and low consistency in their profile (blue)

high consistency tend to receive recommendations that are consistent too. Increasing equality might require playing with the lists recommended to learners that suffer from a low consistency even in their profile.

Optimizing for Equality of Learning Opportunities

With the observations made so far, we conjecture that re-ranking each list of recommendations to maximize the considered principles will lead to higher consistency, and, consequently, to higher equality. Therefore, in this section, we describe, evaluate, and discuss the approach proposed in this paper to favor consistency of the principles (Section 3.2) in recommendations (Fig. 9).

The Post-Processing Approach Proposed

To meet the principles pursued by the platform for each learner and optimizing for equality of opportunities across learners, we introduce a recommendation procedure that seeks to maximize the consistency formalized in (3).

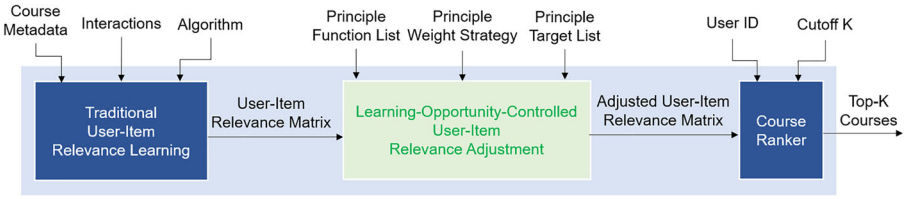


Fig. 9 Support Framework. First, given interactions and metadata, a recommendation algorithm computes a user-item relevance matrix. Then, given a user-item relevance matrix, a list of principle functions, a principle weight strategy, and a list of principle targets, our approach returns a user-item relevance matrix that meets the input principles. Finally, a ranking step, given the adjusted user-item relevance, outputs the recommended list

Given that it is generally hard to build the equity-enhancing mechanisms into the main recommendation algorithm, we propose to re-arrange the recommended lists returned by a recommender system, a common practice known as re-ranking (Potey & Sinha, 2017). On the one hand, this strategy might be limited in its impact, since reordering a small set of recommendations might have a less profound effect than building equality into the recommendation selected by the recommender system from the extensive pool of possibilities. On the other hand, it has several advantages, such as that it can be applied to the output of any recommender system and can be easily extended to include any novel principle. Therefore, we assume that the notion of equal recommended learning opportunities might be enhanced by trying to balance out desirable properties of course recommendations, and that re-ranking the courses originally suggested (and optimized for personalization) by the recommender system can be a feasible strategy for achieving this objective. The re-ranking should operate such that the modified lists recommended to learners meet desirable principles of course recommendations equally across learners.

For each learner $u \in U$, our goal is to determine an optimal set $\mathcal{I}^*$ of k courses to be recommended to u, so that the principles pursued by the platform are met while preserving accuracy (i.e., the extent to which the recommended items are among those included in the test set for that learner, meaning that the recommender system predicts well the future interests of the learner). To this end, we capitalize on a *maximum marginal relevance* (Carbonell & Goldstein, 1998) approach, with (3) as the support metric. In other words, we aim to find the set of courses $\mathcal{I}^*$ to recommended to the learner u such that those courses have high relevance for the learner u ($\widetilde{R}_{ui}$: the relevance predicted by the recommender on the course i for learner u) and their addition to the recommended list brings the highest increase in the consistency level across the principles ($Consistency(p_u, q_{\mathcal{I}}|w)$: the extent to which the degrees p_u for all principles targeted by the platform and the actual degrees $q_{\mathcal{I}}$ these principles are met by the recommended list agree with each other). Let us consider an example where we aim to recommend $k = 10$ courses to a given learner u. For each position p of the ranking, for each course, we compute the weighted sum between (i) the relevance of that course for the learner u and (ii) the consistency the recommended list to u would achieve if we include that course in the list of recommendations. The weight λ assigned to the consistency term allows us to define how important the consistency is concerning the relevance of that course for the learner (i.e., the degree that course meets the individual interests of that learner). Once we compute this weighted score for all courses, we find the course that achieves the highest weighted score, and we add it to the recommendations to u at position p. The same procedure is repeated similarly for the other positions till k.

The set $\mathcal{I}^*$ is obtained by solving the following optimization problem:

$$\mathcal{I}^*(u|k, w) = \underset{\mathcal{I} \subset I, |\mathcal{I}|=k}{\operatorname{argmax}} \; (1-\lambda) \sum_{i \in \mathcal{I}} \widetilde{R}_{ui} + \lambda \, Consistency(p_u, q_{\mathcal{I}}|w), \qquad (6)$$

where $q_{\mathcal{I}}$ is q when the top-k list includes items $\mathcal{I}$, and $\lambda \in [0, 1]$ is a parameter that expresses the trade-off between accuracy and learning opportunity consistency. With $\lambda = 0$, we yield the output of the recommender, not taking consistency optimization

into account. Conversely, with $\lambda = 1$, the output of the recommender is discarded, and we focus only on maximizing consistency.

This greedy approach yields an ordered list of resources, and the resulting list at each step is $(1-1/e)$ optimal among the lists of equal size. The proof of the optimality of the proposed approach is provided in Appendix B. This property fits with the real world, where learners may initially see only the first k recommendations, and the remaining items may become visible after scrolling. Our approach also allows controlling more than one learning opportunity principle in the ranked lists, with no constraints on the size of C.

Evaluation Scenario and Experimental Results

In this section, we assess the impact of controlling consistency and equality of learning opportunities across learners after applying our procedure to pursue the platform's principles (i.e., maximizing all the principle indicators). It is important to note that we considered the same setup described for the exploratory analysis, including the same datasets ("Data"), protocols ("Recommendation Algorithms and Protocols"), and metrics ("Problem Formulation"), to answer four key research questions:

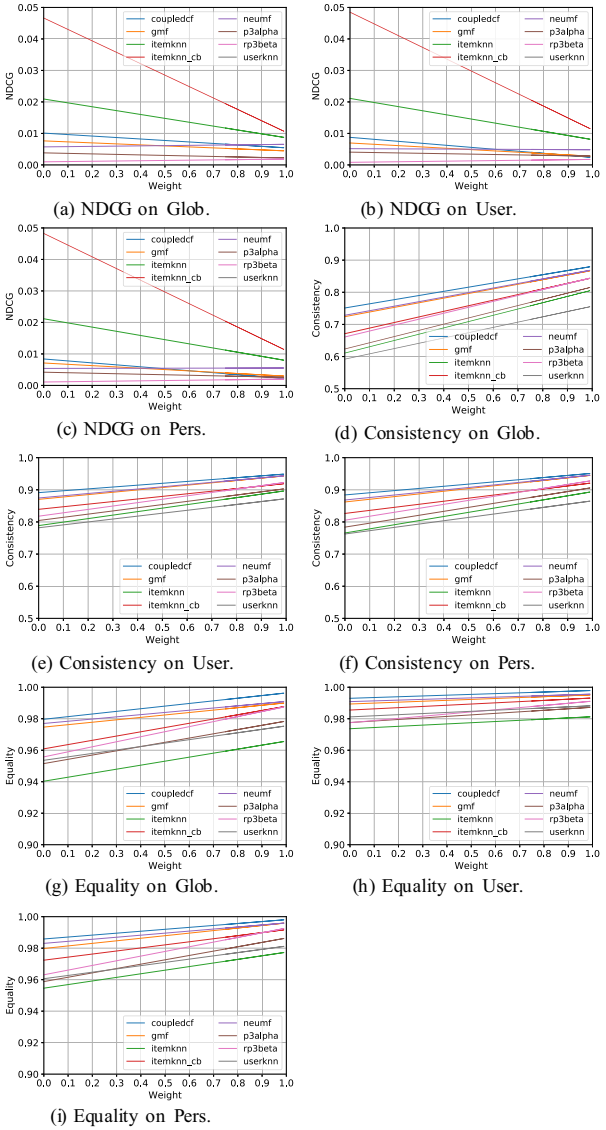
RQ5.2.1 Which weight setup achieved the best accuracy-equality trade-off?
RQ5.2.2 Which principles have experienced the largest gain in consistency?
RQ5.2.3 Which is the influence of the original relevance score distribution?
RQ5.2.4 How do the recommended lists differ, before and after our approach?

Influence of Weight Setup (RQ5.2.1) In this subsection, to answer the first research question, we explore the extent to which each principles' weight setup meets the accuracy-equality trade-off. Given that the consistency achieved with the originally recommended lists is different across principles and learners, different weight setups might lead to distinct levels of the mentioned trade-off. Finding the weight setup that results in the best trade-off is therefore of primary importance to increase equality. We run experiments to assess (i) the influence of our procedure and the weight-based strategy on accuracy, consistency, and equality, and (ii) the relation between a loss in accuracy and a gain in consistency and equality while applying our procedure. To this end, we envisioned three approaches of principle weight assignment:

- `Glob` assigns the same weight to all the principles, for all users. This method would not account for the level of consistency the recommended list to a given user already achieved and will treat all the principles equally.
- `User` assigns, to a principle, a weight proportional to the consistency gap for that principle concerning the target of the platform, computed during the exploratory analysis. The consistency gap for a principle has been obtained by averaging the individual consistency gaps across users.
- `Pers`, given a user, assigns the weight for a principle by considering only their (individual) consistency gap for that principle. Thus, different weights are used along with the user population.

For each model, we run an instance of our re-ranking procedure for each weight assignment strategy, assigning to λ a value in [0.01, 0.25, 0.50, 0.75, 0.99].

The results related to NDCG, consistency, and equality are shown in Fig. 10. Specifically, top-row plots on NDCG highlighted that ItemKNN and ItemKNN-CB experienced the largest loss in NDCG at increasing λ. The rest of the algorithms



(a) NDCG on Glob. (b) NDCG on User.

(c) NDCG on Pers. (d) Consistency on Glob.

(e) Consistency on User. (f) Consistency on Pers.

(g) Equality on Glob. (h) Equality on User.

(i) Equality on Pers.

Fig. 10 Controlled Performance. Normalized Discounted Cumulative Gain (NDCG), consistency, and equality achieved by our procedure under Glob, User, and Pers weight assignment strategies. For each algorithm, our approach has been applied at various λ

showed a more stable pattern on NDCG, even though the NDCG absolute value is significantly lower for the one achieved by ItemKNN and ItemKNN-CB. Throughout the weight assignment strategy, we did not observe a significant difference for the same algorithm over the three strategies. On the other hand, the weight assignment strategy has a notorious role in consistency and equality (middle and bottom rows). Specifically, `User` and `Pers` weight setups made it possible to achieve higher consistency and equality than `Glob`. We can also observe that all the algorithms brought the same degree of improvement in consistency while varying λ.

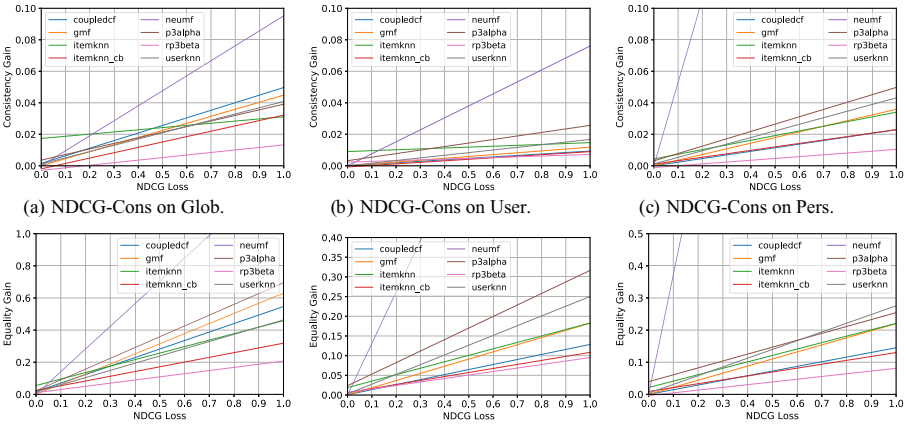
Interestingly, by looking at equality scores, two patterns of improvement were observed. Specifically, the algorithms from the graph-based, content-based, and hybrid families showed a larger improvement at each value of λ than the other families. The following observation can be drawn:

Observation 6. *The considered weight assignment strategies do not differ in terms of accuracy loss. However, User and Pers lead to consistency and equality values higher than Glob, at the same λ. This property means that higher equality of recommended learning opportunities can be achieved by considering the consistency gaps experienced by the individual learner for each principle.*

To have a more detailed picture, we analyzed the connection between a loss in NDCG and a gain in consistency and equality. This aspect plays a key role in a real-world context. While it is the responsibility of scientists to bring forth the discussion about metrics, and possibly to design algorithms to optimize them by turning parameters, it is ultimately up to the stakeholders[12] (e.g., teachers, instructional designers, platform owners), depending on the targeted educational domain, to select the trade-offs most suitable for their context. Therefore, this aspect would support a decision regarding the value of λ to set up to achieve the desired trade-off. Figure 11 plots the gain of consistency (top row) and equality (bottom row) resulting from the degree of NDCG loss. It should be noted that the gain in consistency (equality) is computed with respect to the original consistency (equality) at $\lambda = 0.01$. We observe that consistency and equality within the same weight strategy show the same behavior on the loss in NDCG. This observation confirms the results of our exploratory analysis, where consistency and equality were directly proportional.

We can conclude that the principles' weight setup has a high impact on the accuracy-equality trade-off. Using user-based weights that represent the average of the individual consistency gaps across learners or individual weights that are personalized for each user lead to higher equality of recommended learning opportunities.

[12] Although this approval lets stakeholders control the different factors impacting on fairness, it still leaves open questions around the intrinsic maturity of the weight settings and the accountability in the decision-making process.



(a) NDCG-Cons on Glob. (b) NDCG-Cons on User. (c) NDCG-Cons on Pers.

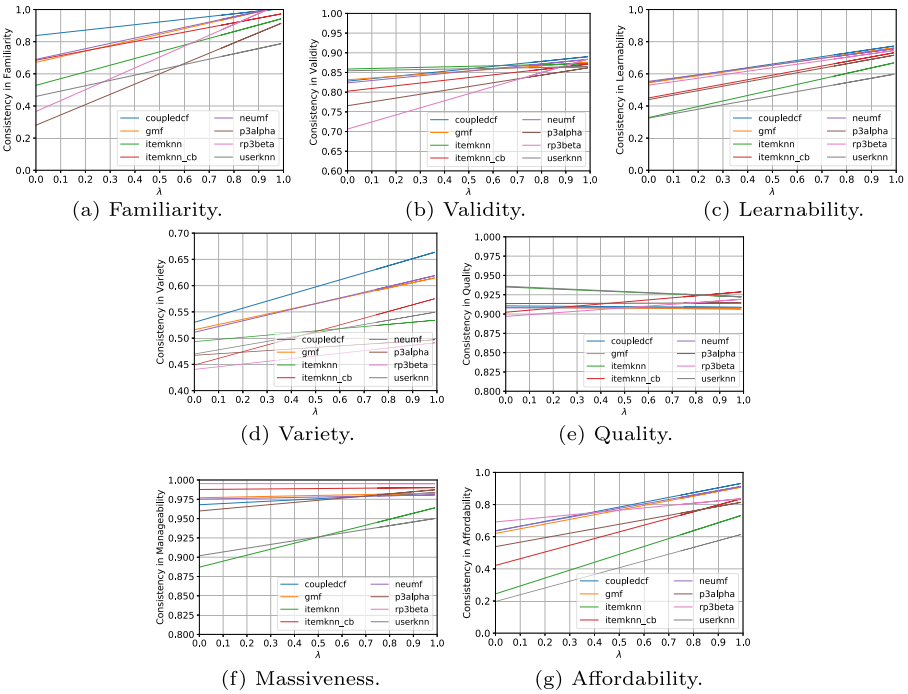
(d) NDCG-Equality on Glob. (e) NDCG-Equality on User. (f) NDCG-Equality on Pers.

Fig. 11 Accuracy-Equality Relation. For each algorithm and weight assignment setup, we computed the gain in equality and consistency that can be achieved at the cost of loosing a certain degree of accuracy

Influence on Each Principle (RQ5.2.2) In this subsection, we answer the second research question, aimed at exploring which principles have experienced the largest gain in consistency, with our approach. For instance, this aspect is important to understand whether our approach will favor those principles that already have high consistency or those principles that suffer from a low consistency. To this end, we run experiments to assess (i) which principles show the largest improvement thanks to the proposed approach, and (ii) what is the impact of the weighting strategy on the consistency of each principle. To answer these questions, for each model, we run an instance of our re-ranking procedure for each weighting strategy, varying $\lambda \in [0.01, 0.25, 0.50, 0.75, 0.99]$. Then, we computed the consistency of each principle achieved by an algorithm, at a given λ, with a given weighting.

Figure 12 reports the impact of our procedure on the considered principle for different algorithms. Overall, it can be observed that our procedure allows us to improve the consistency for all the principles, except Quality (see Fig. 12e). This principle exhibited two main patterns based on the algorithms: quality increased for ItemKNN and RP3Beta, while it decreased for the other algorithms. Interestingly, adding course metadata information into the algorithm (ItemKNN-CB concerning ItemKNN) changes the trend in quality. Furthermore, our approach made it possible to improve Familiarity, Variety, and Affordability, all of which achieved low consistency scores in the exploratory analysis. It follows that the value of λ can be fine-tuned to reach the desired level for a given principle. Another observation can be drawn.

Observation 7. *Controlling learning opportunity results in higher consistency for all principles, except for Quality and Massiveness that maintain stable consistency scores. Quality may decrease in some cases with collaborative filtering.*



(a) Familiarity. (b) Validity. (c) Learnability.

(d) Variety. (e) Quality.

(f) Massiveness. (g) Affordability.

Fig. 12 Controlled Consistency. Consistency per principle achieved by our procedure under the Glob weight assignment strategies at various λ

We can conclude that our approach tends to increase the consistency of those principles that suffer from a lower consistency in the original recommendations. The impact on principles with an initial high consistency is negligible.

Influence of Relevance Score Distribution (RQ5.2.3) Having observed that the improvement in consistency greatly varies among algorithms, we conjecture that the distribution of relevance scores returned by the original algorithm may influence the feasibility of our approach. Therefore, in this subsection, we aim to answer the third research question on the influence of the original relevance score distribution in the results. This aspect might inform us on the characteristics of the relevance scores and, by extension, of the recommender system that might work better with our approach. Hence, Fig. 13 shows the density of relevance scores along the range $[-1, 1]$. We notice that GMF, NeuMF, and ItemKNN-CB produced relevance scores with a high density around zero. Therefore, in our approach, the relevance part may be dominated by the consistency part, regardless of the applied λ. Consequently, relevance could have a drastic drop even for low λ values, making it harder to find a good trade-off between accuracy and consistency. This behavior is confirmed by the results previously reported in Fig. 11. The NDCG loss compared to the Consistency gain is higher for GMF, NeuMF, and ItemKNN-CB. Given that the relevance scores distribution is

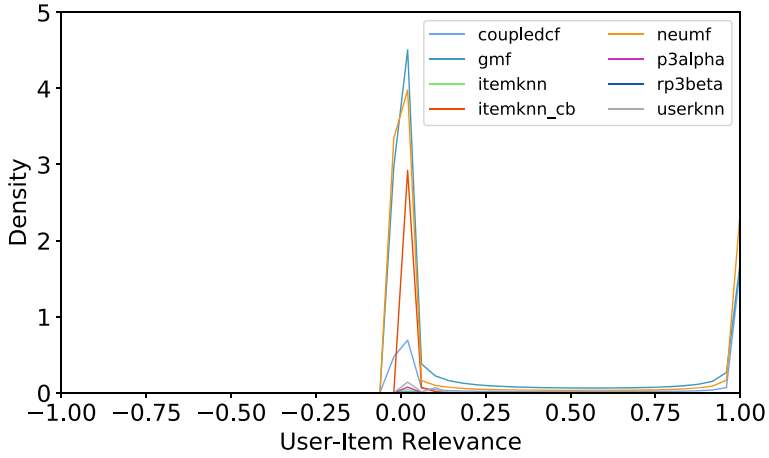


Fig. 13 Relevance Score Distribution. For each algorithm, we compute the density of the user-item relevance scores computed by the original version of the recommender

highly dense, even a small improvement in consistency may completely overturn the list of recommendations.

We can conclude that the density of the relevance score distribution influences the trade-off between accuracy and consistency, after applying our approach. The higher the density is, the higher the drop in accuracy is.

Qualitative Inspection (RQ5.2.4) For the fourth research question, we investigate the extent to which the recommended lists of courses differ, before and after applying our approach. Exploring this perspective is important given that it refers to the concrete differences in recommendations that will be experienced by learners in the real world. The targeted principles and the corresponding consistency and equality metrics directly monitor properties of the recommender lists and the experiments showed that our approach leads potentially[13] to more consistent and equal recommended learning opportunities. However, it may also be interesting to inspect some recommended lists resulting from a traditional recommendation algorithm and how they change after our approach, closing the circle for the problem that motivated this study and assessing the practical end-to-end impact of the proposed approach.

Table 2 shows how the list of courses recommended to an example learner changes after applying our procedure. First, regarding affordability, we observe that the re-ranked list offers a broader range of opportunities in terms of fees, even among

[13]Given that it is based on assumptions and technical implementation, our approach remains to be further evaluated from a human-centered point of view (e.g., on the learners' perceptions in subjective experiences in both academic and life-long outcomes). There is also the need to validate the value of λ in online settings with learners, which remains in the domain of subjective decision-making by course designers and engineers.

Table 2 Impact on Recommendation. Top-10 recommendations provided to an example learner by the traditional recommender system based on ItemKNN-CB (top) and the top-10 recommendations resulting from our approach

	Item ID	Category	Fee	Level	Update	Learners	Asset
Before							
0	17175	development	99.99	all	2017-08-01	778	V+A
1	8518	development	199.99	all	2020-03-02	23386	V+A+E
2	17772	academics	199.99	all	2019-07-30	6501	V+A
3	18689	development	199.99	all	2019-08-11	10588	V+A
4	9364	development	99.99	all	2016-03-01	3825	V+A+E
5	17735	business	199.99	all	2020-03-13	19615	V
6	7932	development	29.99	beg	2015-11-06	4661	V+E
7	13191	development	199.99	all	2019-02-20	133490	V+A
8	15861	development	99.99	all	2020-02-24	24705	V+A
9	5676	development	199.99	all	2020-02-15	10638	V+A
After							
0	4878	development	0.00	beg	2018-07-23	414169	V+A+E
1	9797	it-and-sw	0.00	int	2019-04-12	5605	V+A
2	17175	development	99.99	all	2017-08-01	778	V+A
3	7932	development	29.99	beg	2015-11-06	4661	V+E
4	9364	development	99.99	all	2016-03-01	3825	V+A+E
5	11275	it-and-sw	19.99	int	2019-08-30	777	V
6	8518	development	199.99	all	2020-03-02	23386	V+A+E
7	3707	development	19.99	beg	2015-02-27	1357	V
8	15861	development	99.99	all	2020-02-24	24705	V+A
9	8944	it-and-sw	19.99	beg	2018-11-10	8282	V+E

courses from the same category. This aspect may enable a learner to receive suggestions that can better fit with their current financial resources. Then, more diverse opportunities were proposed in terms of instructional level and asset types, in line with the targets pursued by the platform. Except for the course with a large class ranked in the first position, our approach leads to courses with smaller and, thus, more manageable classes. However, this comes at the price of a slight loss of validity and category diversity. This happened because the learner mostly interacted with "development" and "it-and-software" courses in the past, so our approach promoted courses aligned with those categories (i.e., increasing familiarity).

While the proposed approach confirmed its feasibility for conveying multiple principles into a recommended list and providing more equal learning opportunities across learners, it should be noted that it is ultimately up to the stakeholders to select principles and trade-offs most suitable for their context.

We can conclude that the impact of our approach is concretely observable in the final recommended lists. Specifically, our approach mainly adds to the recommenda-

tions a subset of courses that were originally ranked below the top-10 for that learner, such that the targeted principles are consistently met.

Discussion

There has been an increasing attention to digitalized educational systems. Hence, online course platforms are promptly becoming an essential tool for providing learners with the most suitable material, meeting their expectations of educational values. Due to the highly subjective and contextual nature of this process, educational platforms need to consider multiple perspectives. Indeed, besides providing a wide range of course filtering options, an increasingly high number of principles for further processing such options is needed to identify the most suitable ones for a learner. In this view, an in-depth understanding of recommendations in online course platforms may reduce the overload of learners, improving consistency and equality of the recommendations.

Though the learners of COCO may not be representative of general learners in the recommender system, our analysis in "Exploratory Analysis" indicates that optimizing recommendation algorithms only for learners' interests may result in undermining other essential properties conveyed by the learning opportunities proposed to them. Ranges of educational recommender systems, such as those provided by Bridges et al. (2018), Rieckmann (2018), and Bhumichitr et al. (2017), can thus capitalize on our definitions, metrics, and procedures as a means for assessing recommendations' consistency. However, the principles proposed in this paper, derived from curriculum design beliefs, would need to be empirically-validated.

A complementary human-centered perspective of the principle design can strongly benefit from our findings, leveraging our initial principles as a good starting point. Nonetheless, this approach might not be sufficient, and no one-size-fits-all set of principles would exist, given their dependence on the context and the involved stakeholders. Despite having a range of limitations in terms of context-sensitivity, learner-centeredness, operationalization, and temporal awareness ("Limitations"), the re-ranking procedure that was proposed and assessed in "Optimizing for Equality of Learning Opportunities" has been shown to improve equality across learners, counteracting potential pitfalls of data-driven educational recommender systems. This aspect becomes of paramount importance in large-scale contexts, especially while reaching out to learners reluctant to the use of data-driven procedures (Herold, 2017). However, our results cannot prove that the differences in measured metrics translate to better educational outcomes and learners' acceptance. Finally, our work embeds views and needs of multiple educational stakeholders into recommended lists (Abdollahpouri et al., 2020).

In the broad discussion on FATE in AIED, we highlight that recommender capabilities are an important component of AIED systems. Moreover, our research contributes to the improvement of our understanding of fairness in the educational recommendation context, by devising ways in which we can address fairness in AIED design. Our study moves a step forward in understanding how equality principles can be operationalized and combined in a formal notion of equal opportunities in educational recommendations. This contribution serves as a foundation to investigate how

learners interpret concepts such as: (1) the fairness of the educational resource selection decisions they make (e.g., how they select courses for their degrees); (ii) what they think about the fairness of the set of resources available to them; (iii) to what extent they view the course selection process as fair; and (iv) how these decisions are influenced by available information about the given courses. The principles and formulations described in our study would be a starting point for this purpose. Therefore, this study permits the research community to derive what questions to ask as part of interviews with learners, what scenarios to explore to elicit their concepts of fairness, and how to process data in the educational platform to monitor and ensure the equality principles. This paper, thus, shapes a blueprint of the decisions and processes to be done, once empirically-validated principles have been defined under the targeted educational scenario. To this end, we provide evidence on what kind of technological support is needed to ensure that learners' course selection decisions lead to greater equality across learners. Nonetheless, there are several issues regarding inequality in educational opportunities that recommender systems could not fix by themselves (e.g., whether certain advanced courses are available at all). Consequently, our study would help better understanding what is known about the role that recommender systems could play in the bigger questions around fairness and equality, grounding the design and implementation of formal notions of equality informed by a deep understanding of how learners view equality.

Limitations

Since our observations varied over algorithms and principles, we identified the main implications and limitations of our study.

- **Limitations of data**. While our results highlight the need to consider equality of recommended learning opportunities while evaluating recommenders, the learners of COCO may not be representative of general learners in an educational platform. Unfortunately, data with enough attributes to look for sophisticated principles is hard to find. As pointed out in "Data", other datasets include few attributes of learners and courses.
- **Limitations of principles**. While our principles shed light on important aspects underlying the ranked courses, they may not be representative of the principles targeted by certain platforms and are based on assumptions derived from a dataset. Thus, limitations arise from different perspectives.
    - **Context dependency**. Because this study does not provide formative or summative results about actual systems, it is thus more theoretical than practical and does establish a framework for work in the context of fairness in educational recommender systems. In traditional scenarios, the operationalization of principles is usually based on textual guidelines, and the translation into numerical indicators (when performed) is subjected to the specificity of the platform. Our measures show how traditional text-based principles could be operationalized and enriched based on the peculiarity of the online context (e.g., manageability).

However, our framework can be adapted to any (number of) principles, based on the pursued goals.

- **Learner-centeredness**. The considered principles do not relate to individuals, but to resources, educational level, number of learners, and so on, referring to courses as a group. Given that learning and teaching deal with changes in individuals (e.g, how each learner reacts to a problem within a course or which asset in which course was valuable for learners), our principles could be enriched to reach this level of modeling, e.g., by adding principles connected with learning outcomes, when large enough datasets will become available to the research community. It remains also to be investigated the extent to which learners philosophically agree with this approach to equality, agree with it in practice, and are willing to accept the downsides. Finally, empirical evidence would be needed to assess whether recommendations from a recommender system and re-ordering thereof, have any effect on how learners behave (i..e, what courses they actually take).
- **Technical operationalization**. Some of the operationalizations have been overly simplified due to the limitations of the data currently available in online course platforms. For instance, the recency of updates is used as a proxy for validity, but there is more to validity than the recency of updates. Similarly, learners' ratings are used as a proxy for quality, but learners' ratings would not always correlate with other measures of quality (e.g., learning outcomes), and class size is used as a proxy for manageability, but other aspects (e.g., number of assistants) are not captured. Finally, the way learnability is operationalized seems to be simplified, and other aspects might be targeted. This opens up to more advanced operationalizations.
- **Temporal influence**. Some of the principles are sensitive to time. For instance, regarding familiarity, sometimes a learner may be looking for something new that broadens their horizon, rather than something familiar; at other times, they may need one more final elective for their primary major, which might mean that the preferred resource would have high familiarity. In other words, it is still not clear that a given learner should be viewed as having a preference for a certain level of familiarity per se. The same learner may, at different times, have different preferences. Similarly, regarding the measure of validity (i.e., recency of the last update), a course on foundational material could have been updated many times in the past and does not benefit from recent updates. Similar observations apply to other principles.

- **Limitations of the ethical constructs**. Given that our methodology has been assessed in an offline setting, the real-world validity of the notion of equality that is presented still needs to be shown. For instance, it should be investigated whether this notion aligns with learner's notion of fairness, whether learners pay attention to all principles when assessing fairness, and whether this notion of equality relates to fairness and ethical principles, especially if enhancing equality

requires to give up a given degree of personalization. For instance, it is worth exploring how fair it is to ask one learner to give up a degree of personalization so that the familiarity target for another learner can be met. These issues will drive future research of equality of recommended learning opportunities.

- **Limitations of algorithms**. Our study involves eight representative algorithms from four families, but other types of algorithms may benefit from our procedure. However, to better focus on the evaluation of our contribution and due to the limitations of the data, we constrained our study to algorithms that are key building blocks of several recommender systems.
- **Limitations of evaluation protocol**. Our results cannot prove that the differences in measured metrics translate to better educational outcomes and learners' acceptance. Further studies with online evaluation are needed to complement these results. However, we conjecture that our results can provide an essential contribution to reach this goal, and offline protocols can be useful to select algorithms prior to an online deployment.
- **Limitations of metrics**. Among the large number of metrics that can be used for evaluating a recommender system, we focus on consistency and equality to better assess our contribution. We also measured NDCG because it maps well to recommendation utility. However, consistency and equality do not consider the position of the courses in a list, which can be important in large-scale recommendation contexts (e.g., online course platforms where tons of courses are provided and having courses at the top of the recommended list is crucial to visibility), as an example. Our study focused on a more general perspective to reach a broader audience.

Conclusions

In this paper, we proposed a novel fairness metric that monitors the equality of learning opportunity across learners in the context of educational recommender systems, according to a novel set of educational principles. Then, we explored the learning opportunities provided by ten state-of-the-art recommender systems in a large-scale online course platform, uncovering systematic inequalities across learners. To counteract this phenomenon, we proposed a post-processing approach that re-ranks the recommended courses originally returned by an algorithm to maximize the equality of recommended learning opportunities while preserving personalization. Finally, we assessed the impact of supporting learners with our approach to accuracy and beyond-accuracy metrics. Based on the results, we can conclude that:

1. Recommendation algorithms tend to produce ranked lists with low equality of recommended learning opportunities across learners, especially when the algorithm uses only user-item interactions as training data.
2. Under our definition of the targeted principles, equality of quality, validity, and manageability are guaranteed by recommenders. Familiarity, affordability, learnability, and variety exhibit strong deviations over algorithms.

3. Optimizing recommendations for consistency concerning a set of principles leads to higher equality of recommended learning opportunities. This effect is remarkable when learner-specific weights are adopted.
4. Controlling learning opportunity results in higher familiarity, variety, and affordability while maintaining stable values for the other principles. However, quality may experience small losses after applying our procedure.
5. The impact of our approach on accuracy and consistency depends on the density of the relevance score distribution of the original recommendation algorithm. The higher the density, the higher the drop in accuracy is.

Future work will embrace our findings to study the degree to which the courses currently attended by learners satisfy the notion of equality, in addition to the courses that are recommended. Moreover, a learner-centered approach will be carried out to investigate what learner's notions are for the fairness of the educational resource selection decisions they make, to fine-tune and adjust our original set of principles. By extension, learner-specific targeted degrees for each principle will be, consequently, elicited and applied. Thanks to its flexibility, the notions and procedures proposed in this study can fit with a plethora of applications within both educational and non-educational contexts. There is also room for considering how additional algorithms respond to evaluation and what internal mechanics contribute to achieving higher consistency and equality. Finally, as real-world applications should consider whether their recommender systems provide consistent and equal learning opportunities across learners, we believe that there will be an increasing amount of research related to applying our study to the educational industry.

With this study, we highlighted that our notions and procedures are quite broad and incorporate elements of societal and ethical importance. It may be inevitable that, as recommender systems move further into education, they will embed strategies like the one we presented.

Appendix A: Mathematical Notation for Targeted Educational Principles

In this appendix, we provide the mathematical formulations associated with the educational principles proposed in "Modeling Recommended Learning Opportunity through Principles". They have been adopted for computing to what extent each principle is achieved for each learner throughout the experiments.

Familiarity Given a course feature $F_1 \in \mathcal{N}$ associated with integer-encoded representation of the category $g \in G$ of a resource, we consider two distributions:

- $x(g|u)$: the distribution over categories G of the set of resources I_u user u interacted with in the past, defined as $x(g|u) = |I_u^g|/|I_u|$;
- $y(g|u)$: the distribution over categories G of the set of learning opportunities $\tilde{I}_u$ recommended to learner u, defined as $y(g|u) = |\tilde{I}_u^g|/|\tilde{I}_u|$;

where I_u^g and $\tilde{I}_u^g$ represent the set of resources belonging to category g the learner u attended and the recommender system proposed, respectively. Then, we define the familiarity of $\tilde{I}_u$ for a learner u as the inverse of the Hellinger distance across $x(G|u)$ and $y(G|u)$. Specifically:

$$c_{\tilde{I}_u}(1) = 1 - H(x(G|u), y(G|u)) \tag{7}$$

where $c_{\tilde{I}_u}(1) = 1$ if x_u and y_u are perfectly balanced, and the highest familiarity is achieved. Conversely, the minimum familiarity 0 is achieved when x_u assigns probability 0 to every event that y_u assigns a positive probability (or vice versa). In the latter situation, the recommender suggests resources opposite with respect to the user's most familiar categories.

Validity Given a course feature $F_2 \in \mathcal{N}$ representing the last time a resource has been updated and the opening time of the platform, denoted as T_o, we define the validity of a set of learning opportunities $\tilde{I}_u$ at the current time T_c as follows:

$$c_{\tilde{I}_u}(2) = 1 - \frac{1}{|\tilde{I}_u|} \sum_{i \in \tilde{I}_u} \frac{T_c - f_{2,i}}{T_c - T_o} \tag{8}$$

where values close to 0 mean that the learning opportunities are obsolete, while values close to 1 correspond to mostly fresh opportunities in $\tilde{I}_u$.

Learnability Given a course feature $F_3 \in \mathcal{N}$ representing the instructional level of a resource, we define the learnability in $\tilde{I}_u$ as:

$$c_{\tilde{I}_u}(3) = 1 - GINI\left(\frac{|\tilde{I}_u^{f_3}|}{|\tilde{I}_u|} \;\forall\; f_3 \in F_3\right) \tag{9}$$

where $c_{\tilde{I}_u}(3)$ is the inverse of Gini inequality index over the representations of all the instructional levels in $\tilde{I}_u$, and $\tilde{I}_u^{f_3}$ is the set of resources in $\tilde{I}_u$ with instructional level f_3. A value of 0 implies large inequality, while high balance is obtained with values close to 1.

Variety Given that each resource $j \in \tilde{I}_u$ is composed from a set of assets L_j and that the asset type of a resource j is denoted by $T_j = (t_l \in T \;:\; \forall l \in L_j)$, we define the variety of the types in $\tilde{I}_u$ as:

$$c_{\tilde{I}_u}(4) = \frac{1}{|\tilde{I}_u|} \sum_{i \in \tilde{I}_u} \frac{|T_i|}{|T|} \tag{10}$$

where values close to 0 mean that the learning opportunities are focused on few asset types, while asset types greatly vary for values close to 1.

Quality Given a learner-resource feedback's matrix R and that the platform allows for ratings between $F5_{min}$ to $F5_{max}$, we define the quality of a set $\tilde{I}_u$ as follows:

$$c_{\tilde{I}_u}(5) = 1 - \frac{1}{|\tilde{I}_u|} \sum_{i \in \tilde{I}_u} \frac{1}{|U_i|} \sum_{u \in U_i} \frac{F5_{max} - R_{u,i}}{F5_{max} - F5_{min}} \tag{11}$$

where values close to 0 mean that the learning opportunities are of low quality, while values close to 1 are measured for high-quality opportunities.

Manageability Given a course feature $F_6 \in \mathcal{N}$ representing the number of enrolled learners in a course and that the platform allows for classes from $F6_{min}$ to $F6_{max}$ learners, we define the manageability in a set of learning opportunities $\tilde{I}_u$ as follows:

$$c_{\tilde{I}_u}(6) = \frac{1}{|\tilde{I}_u|} \sum_{i \in \tilde{I}_u} \frac{F6_{max} - f_{6,i}}{F6_{max} - F6_{min}} \tag{12}$$

where values close to 1 mean that the learning opportunities include small classes, while values close to 0 refer to large classes.

Affordability Given a course feature $F_7 \in \mathcal{R}$ representing the course enrollment fee and that the platform allows for courses with a cost between $F7_{min}$ and $F7_{max}$, we define the affordability of a set of learning opportunities $\tilde{I}_u$ as follows:

$$c_{\tilde{I}_u}(7) = \frac{1}{|\tilde{I}_u|} \sum_{i \in \tilde{I}_u} \frac{F7_{max} - f_{7,i}}{F7_{max} - F7_{min}} \tag{13}$$

where values close to 0 mean that the learning opportunities are highly expensive, while values close to 1 correspond to free-of-charge learning opportunities in $\tilde{I}_u$.

Appendix B: Optimality Proof for the Proposed Post-Processing Approach

The combinatorial maximization problem in (6) may be efficiently approximated with a greedy approach with $(1 - 1/e)$ optimality if the objective function of the maximization is submodular. This statement has been proved in the following demonstration.

Theorem 1 *Let* $Consistency(p, q|w) = 1 - w\|p - q\|_{|C|}^{|C|}$, *with* $|C|F > 0$ *and* $w_i \geq 0 \; \forall i \in \{0, \cdots, |C|\}$, *then for any* $\lambda \in [0, 1]$ *the function in (6),*

$$f(\mathcal{I}|w) = (1 - \lambda) \sum_{i \in \mathcal{I}} \widetilde{R}_{ui} + \lambda \, Consistency(p_u, q_{\mathcal{I}}|w),$$

is submodular. ∘

Proof First, since $\widetilde{R}_{ui} > 0$, it follows that $f_1(\mathcal{I}|w) = \sum_{i \in \mathcal{I}} \widetilde{R}_{ui}$ is a modular function (i.e., hence, also submodular), because it is a sum of positive quantities.

Second,

$$f_2(\mathcal{I}|w) = Consistency(p_u, q_\mathcal{I}) = w\|p_u - q_\mathcal{I}\|_{|C|}^{|C|}$$
$$= \sum_{i=1}^{k} w_i \ |[p_u]_i - [q_\mathcal{I}]_i|^{|C|} = \sum_{i=1}^{k} x_i,$$

where $x_i = w_i \ |[p_u]_i - [q_\mathcal{I}]_i|^{|C|} > 0$. Again, f_2 is modular because it is a sum of positive quantities. Since $f(\mathcal{I}|w) = (1-\lambda) f_1(\mathcal{I}|w) + \lambda f_2(\mathcal{I}|w)$, and the convex combination of submodular functions is submodular, f is submodular. □

Acknowledgments This work has been partially supported by the Sardinian Regional Government, POR FESR 2014-2020 - Axis 1, Action 1.1.3, under the project "SPRINT" (D.D. n. 2017 REA, 26/11/2018, CUP F21G18000240009), and by the Agència per a la Competitivat de l'Empresa, ACCIÓ, under the project "Fair and Explainable Artificial Intelligence (FX-AI)". Furthermore, this work was supported in part by FCT project POCI-01-0145-FEDER-031411-HARMONY.

Funding Open Access funding provided by EPFL Lausanne.

References

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Complementing educational recommender systems with open learner models. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 360–365).

Abdollahpouri, H., Adomavicius, G., Burke, R., Jannach, D., Kamishima, T., Krasnodebski, J., & Pizzato, L. (2020). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, *30*(1), 127–158.

Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. *International Educational Data Mining Society*.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. In *Thirty-first conference on neural information processing systems, NIPS* (p. 2017).

Beattie, I. R., & Thiele, M. (2016). Connecting in class? college class size and inequality in academic social capital. *The Journal of Higher Education*, *87*(3), 332–362.

Beutel, A., Chen, J., Doshi, T., Qian, H., Li, W., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., & Goodrow, C. (2019). Fairness in recommendation ranking through pairwise comparisons. In *International conference on knowledge discovery & data mining, KDD* (pp. 2212–2220). ACM.

Bhumichitr, K., Channarukul, S., Saejiem, N., Jiamthapthaksin, R., & Nongpong, K. (2017). Recommender systems for university elective course recommendation. In *2017 14th international joint conference on computer science and software engineering (JCSSE)* (pp. 1–5). IEEE.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *41St international ACM conference on research & development in information retrieval, SIGIR* (pp. 405–414).

Boratto, L., Fenu, G., & Marras, M. (2019). The effect of algorithmic bias on recommender systems for massive open online courses. In *European conference on information retrieval, ECIR* (pp. 457–472). Springer.

Boxuan, M.A., Taniguchi, Y., & Konomi, S. (2020). Course recommendation for university environments. In *Thirteenth international conference on educational data mining (EDM 2020)*.

Bridges, C., Jared, J., Weissmann, J., Montanez-Garay, A., Spencer, J., & Brinton, C. G (2018). Course recommendation as graphical analysis. In *2018 52nd annual conference on information sciences and systems, CISS* (pp. 1–6). IEEE.

Buchholz, S., Skopek, J., Zielonka, M., Ditton, H., Wohlkinger, F., & Schier, A. (2016). Secondary school differentiation and inequality of educational opportunity in germany. In *Models of secondary education and social inequality*. Edward Elgar Publishing.

Bulathwela, S., Yilmaz, E., & Shawe-Taylor, J. (2019). Towards automatic, scalable quality assurance in open education. In *Workshop on AI and the united nations SDGs at international joint conference on artificial intelligence*.

Bulger, M. (2016). Personalized learning: The conversations we're not having. *Data and Society, 22*(1).

Byun, S., & Park, H. (2017). When different types of education matter: Effectively maintained inequality of educational opportunity in Korea. *American Behavioral Scientist*, *61*(1), 94–113.

Campos, P. G., Díez, F., & Cantador, I. (2014). Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, *24*(1-2), 67–119.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *ACM Conference on research and development in information retrieval, SIGIR* (pp. 335–336).

Chan, T.-W., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., Patton, C., Cherniavsky, J. C., Pea, R. D., Norris, C., Soloway, E., Balacheff, N., Scardamalia, M., Dillenbourg, P., Looi, C.-K., Milrad, M., & Hoppe, H. U. (2006). One-to-one technology-enhanced learning: an opportunity for global research collaboration. *Research and Practice in Technology Enhanced Learning*, *1*(1), 3–29.

Chanaa, A., & Faddouli, N.-E. E. (2020). Predicting learners need for recommendation using dynamic graph-based knowledge tracing. In *International conference on artificial intelligence in education* (pp. 49–53). Springer.

Chau, H., Barria-Pineda, J., & Brusilovsky, P. (2018). Learning content recommender system for instructors of programming courses. In *International conference on artificial intelligence in education* (pp. 47–51). Springer.

Chen, Z., & Demmans, C. (2020). Epp. Csclrec: Personalized recommendation of forum posts to support socio-collaborative learning. In *Thirteenth international conference on educational data mining (EDM 2020)* (pp. 364–373).

Conaway, W., & Zorn-Arnold, B. (2016). The keys to online learning for adults. *Distance Learning Issue*, *13*, 1.

Cooper, C., Lee, S. H., Radzik, T., & Siantos, Y. (2014). Random walks in recommender systems: exact computation and simulations. In *23Rd international conference on World Wide Web, WWW* (pp. 811–816).

Dai, Y., Asano, Y., & Yoshikawa, M. (2016). Course content analysis: An initiative step toward learning object recommendation systems for mooc learners. *International Educational Data Mining Society*.

Darwin, S. (2017). What contemporary work are student ratings actually doing in higher education? *Studies in Educational Evaluation*, *54*, 13–21.

Dessì, D., Fenu, G., Marras, M., & Reforgiato Recupero, D. (2018). Coco: Semantic-enriched collection of online courses at scale with experimental use cases. In *World conference on information systems and technologies, worldcist* (pp. 1386–1396). Springer.

Doroudi, S., & Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 335–339).

Drachsler, H., Hoel, T., Scheffel, M., Kismihók, G., Berg, A., Ferguson, R., Chen, W., Cooper, A., & Manderveld, J. (2015). Ethical and privacy issues in the application of learning analytics. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 390–391).

Druzhinina, M., Belkova, N., Donchenko, E., Liu, F., & Morozova, O. (2018). Curriculum design in professional education: Theory and practice. In *SHS Web of conferences*, (Vol. 50 p. 01046). EDP sciences.

Eagle, M., Corbett, A., Stamper, J., & Mclaren, B. (2018). Predicting individualized learner models across tutor lessons. *International Educational Data Mining Society*.

Esteban, A., Zafra, A., & Romero, C. (2018). A hybrid multi-criteria approach using a genetic algorithm for recommending courses to university students. *International Educational Data Mining Society*.

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding dropouts in moocs. In *International conference on artificial intelligence*, (Vol. 33 pp. 517–524). AAAI.

Fernández-Mellizo, M., & Martínez-García, J. S. (2017). Inequality of educational opportunities: School failure trends in Spain (1977–2012). *International Studies in Sociology of Education*, *26*(3), 267–287.

Fujihara, S., & Ishida, H. (2016). The absolute and relative values of education and the inequality of educational opportunity: Trends in access to education in postwar japan. *Research in Social Stratification and Mobility*, *43*, 25–37.

Girvan, C. (2018). What is a virtual world? definition and classification. *Educational Technology Research and Development*, *66*(5), 1087–1100.

Golley, J., & Kong, S. T. (2018). Inequality of opportunity in china's educational outcomes. *China Economic Review*, *51*, 116–128.

Gómez-Rey, P., Barbera, E., & Fernández-Navarro, F. (2016). Measuring teachers and learners' perceptions of the quality of their online learning experience. *Distance Education*, *37*(2), 146–163.

Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

Hansen, J. D., & Reich, J. (2015). Democratizing education? examining access and usage patterns in massive open online courses. *Science*, *350*(6265), 1245–1248.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. (2017). Neural collaborative filtering. In *26Th international conference on World Wide Web, WWW* (pp. 173–182).

Herold, B. (2017). The case (s) against personalized learning. *Education Week*, *37*(12), 4–5.

Holmes, W., Iniesto, F., Sharples, M., & Scanlon, E. (2019). Ethics in aied: Who cares? an ec-tel workshop.

Holstein, K., & Doroudi, S. (2019). Fairness and equity in learning analytics systems (fairlak). In *Companion proceedings of the ninth international learning analytics & knowledge conference (LAK 2019)*.

Holstein, K., McLaren, B. M., & Aleven, V. (2019a). Designing for complementarity: Teacher and student needs for orchestration support in ai-enhanced classrooms. In *International conference on artificial intelligence in education* (pp. 157–171). Springer.

Holstein, K., Vaughan, J. W., Daumé III, H., Dudik, M., & Wallach, H. (2019b). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1–16).

Hu, Q., & Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. In *Proceedings of The 13th international conference on Educational Data Mining (EDM 2020)* (pp. 431–437).

Jacobsen, A., & Spanakis, G. (2019). It's a match! reciprocal recommender system for graduating students and jobs. ERIC.

Joyner, D. A., Goel, A. K., & Isbell, C. (2016). The unexpected pedagogical benefits of making higher education accessible. In *Proceedings of the third (2016) ACM Conference on Learning@ Scale* (pp. 117–120).

Khanal, S. S., Prasad, P.W.C., Alsadoon, A., & Maag, A. (2019). A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 1–30.

Kulkarni, P. V., Rai, S., & Kale, R. (2020). Recommender system in elearning: A survey. In *International conference on computational science and applications, ICCSA* (pp. 119–126). Springer.

Kumar, S., Martin, F., Budhrani, K., & Ritzhaupt, A. (2019). Award-winning faculty online teaching practices: Elements of award-winning courses. *Online Learning*, *23*(4), 160–180.

Labarthe, H., Bouchet, F., Bachelet, R., & Yacef, K. (2016). Does a peer recommender foster students' engagement in moocs? International Educational Data Mining Society.

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019a). ifair: Learning individually fair data representations for algorithmic decision making. In *IEEE 35th International conference on data engineering, ICDE* (pp. 1334–1345). IEEE.

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019b). Operationalizing individual fairness with pairwise fair representations. *The VLDB Endowment*, *13*(4), 506–518.

Lin, J., Sun, G., Shen, J., Pritchard, D., Cui, T., Xu, D., Li, L., Beydoun, G., & Chen, S. (2020). Deep-cross-attention recommendation model for knowledge sharing micro learning service. In *International conference on artificial intelligence in education* (pp. 168–173). Springer.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook* (pp. 73–105). Springer.

Lowenthal, P. R., Nyland, R., Jung, E., Dunlap, J. C., & Kepka, J. (2019). Does class size matter? an exploration into faculty perceptions of teaching high-enrollment online courses. *American Journal of Distance Education*, *33*(3), 152–168.

Mao, Y. (2019). One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W (2019). Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 444–460).

Meyer, K. (2016). Why should we demand equality of educational opportunity? *Theory and Research in Education*, *14*(3), 333–347.

Mi, F., & Faltings, B. (2017). Adaptive sequential recommendation for discussion forums on moocs using context trees. In *Proceedings of the 10th international conference on educational data mining, number CONF*.

Mohapatra, S., & Mohanty, R. (2017). Adopting moocs for afforable quality education. *Education and Information Technologies*, *22*(5), 2027–2053.

Morsomme, R., & Alferez, S. V. (2019). Content-based course recommender system for liberal arts education. International Educational Data Mining Society.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501.

Papathoma, T., Ferguson, R., Iniesto, F., Rets, I., Vogiatzis, D., & Murphy, V. (2020). Guidance on how learning at scale can be made more accessible. In *Proceedings of the seventh ACM conference on learning@ Scale* (pp. 289–292).

Pardos, Z. A., & Jiang, W. (2020). Designing for serendipity in a university course recommendation system. In *International conference on learning analytics & knowledge, LAK* (pp. 350–359).

Paudel, B., Christoffel, F., Newell, C., & Bernstein, A. (2016). Updatable, accurate, diverse, and scalable recommendations for interactive applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(1), 1–34.

Pinkwart, N. (2016). Another 25 years of aied? challenges and opportunities for intelligent educational technologies of the future. *International Journal of Artificial Intelligence in Education*, *26*(2), 771–783.

Polyzou, A., Nikolakopoulos, A. N., & Karypis, G. (2019). Scholars walk: A markov chain framework for course recommendation. *International Educational Data Mining Society*.

Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In *Artificial intelligence and inclusive education* (pp. 39–59). Springer.

Potey, M. A., & Sinha, P. K. (2017). Personalization approaches for ranking: A review and research experiments. *International Journal of Information Retrieval Research IJIRR*, *7*(1), 1–16.

Potts, B. A., Khosravi, H., Reidsema, C., Bakharia, A., Belonogoff, M., & Fleming, M. (2018). Reciprocal peer recommendation for learning purposes. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 226–235).

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in moocs. In *ACM International conference on web search and data mining, WSDM* (pp. 93–102).

Qiu, X., & Lo, Y. Y. (2017). Content familiarity, task repetition and chinese efl learners' engagement in second language use. *Language Teaching Research*, *21*(6), 681–698.

Ramos, G., Boratto, L., & Caleiro, C. (2020). On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. *Information Processing & Management*, *57*(2), 102058.

Rastegarpanah, B., Gummadi, K. P., & Crovella, M. (2019). Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *International conference on web search and data mining, WSDM* (pp. 231–239). ACM.

Ren, Z., Ning, X., Lan, A. S., & Rangwala, H. (2019). Grade prediction based on cumulative knowledge and co-taken courses. International Educational Data Mining Society.

Rieckmann, M. (2018). Learning to transform the world: key competencies in education for sustainable development. *Issues and trends in education for sustainable dev*39.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *10Th international conference on world wide web, WWW* (pp. 285–295).

Sclater, N., & Bailey, P. (2015). Code of practice for learning analytics.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68).

Shields, L., Newman, A., & Satz, D. (2017). Equality of educational opportunity.

Shum, S. B. (2018). Transitioning education's knowledge infrastructure: Shaping design or shouting from the touchline? In *Proceedings of International Conference of the Learning Sciences ICLS*.

Singh, A., & Joachims, T. (2019). Policy learning for fairness in ranking. In *Advances in neural information processing systems* (pp. 5427–5437).

Steck, H. (2018). Calibrated recommendations. In *12th ACM conference on recommender systems, recsys* (pp. 154–162).

Talla, M. (2012). *Curriculum development: Perspectives, principles and issues*. India: Pearson Education.

Thaker, K., Zhang, L., He, D., & Brusilovsky, P. (2020). Recommending remedial readings using student knowledge state. In *13th international conference on educational data mining* (pp. 233–244).

Tsai, Y.-S., & Gasevic, D. (2017). Learning analytics in higher education—challenges and policies: a review of eight learning analytics policies. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 233–242).

Vygotsky, L. S. (1978). Mind in society the development of higher psicologycal processes.

Wang, S., Wu, H., Ji, H. K., & Andersen, E. (2019). Adaptive learning material recommendation in online language education. In *International conference on artificial intelligence in education* (pp. 298–302). Springer.

Williamson, B. (2017). Decoding classdojo: psycho-policy, social-emotional learning and persuasive educational technologies. *Learning, Media and Technology*, *42*(4), 440–453.

Xie, I., & Joo, S. (2009). Selection of information sources Accessibility of and familiarity with sources, and types of tasks. *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1–18.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Annual conference on neural information processing systems, NIPS* (pp. 2921–2930).

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: evaluating different sources of student data. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*.

Zhang, J., Hao, B., Bo, C., Li, C., Chen, H., & Sun, J. (2019). Hierarchical reinforcement learning for course recommendation in moocs. In *International conference on artificial intelligence, AAAI*, (Vol. 33 pp. 435–442).

Zhang, Q., Cao, L., Zhu, C., Li, Z., & Sun, J. (2018). Coupledcf: Learning explicit and implicit user-item couplings in recommendation for deep collaborative filtering. In *International joint conference on artificial intelligence, IJCAI*.

Zhu, Z., Hu, X., & Caverlee, J. (2018). Fairness-aware tensor-based recommendation. In *International conference on information and knowledge management, CIKM* (pp. 1153–1162). ACM.

Affiliations

Mirko Marras[1] · Ludovico Boratto[2] · Guilherme Ramos[3] · Gianni Fenu[2]

Ludovico Boratto
ludovico.boratto@acm.org

Guilherme Ramos
gramos@fe.up.pt

Gianni Fenu
fenu@unica.it

1 EPFL, Lausanne, Switzerland

2 University of Cagliari, Cagliari, Italy

3 University of Porto, Porto, Portugal