THIS DOCUMENT CONTAINS THE AUTHOR'S ACCEPTED MANUSCRIPT. THE PUBLISHED MANUSCRIPT – ACCEPTED FOR PUBLICATION IN *APPLIED MEASUREMENT IN EDUCATION – IS AVAIABLE WITH DOI: https://doi.org/10.1080/08957347.2022.2034824*.

Published on April 6, 2022:

Personalized Online Learning, Test Fairness, and Educational Measurement: Considering Differential Content Exposure Prior to a High Stakes End of Course Exam

Daniel Katz
University of California, Santa Barbara

Anne Corinne Huggins-Manley & Walter Leite
University of Florida

Corresponding author: Daniel Katz, dkatz@ucsb.edu; Gevirtz Graduate School of Education

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Katz, D., Huggins-Manley, A. C., & Leite, W. (2022). Personalized Online Learning, Test Fairness, and Educational Measurement: Considering Differential Content Exposure Prior to a High Stakes End of Course Exam. *Applied Measurement in Education*, *35*(1), 1-16.

Abstract

According to the Standards for Educational and Psychological Testing (2014), one aspect of test fairness concerns examinees having comparable opportunities to learn prior to taking tests. Meanwhile, many researchers are developing platforms enhanced by artificial intelligence (AI) that can personalize curriculum to individual student needs. This leads to a larger overarching question: When personalized learning leads to students having differential exposure to curriculum throughout the K-12 school year, how might this affect test fairness with respect to summative, end-of-year highstakes tests? As a first step, we traced the differences in content exposure associated with personalized learning and more traditional learning paths. To better understand the implications of differences in content coverage, we conducted a simulation study to evaluate the degree to which curriculum exposure varied across students in a particular AI-enhanced learning platform for Algebra instruction with high-school students. Results indicate that AI-enhanced personalized learning may pose threats to test fairness as opportunity-to-learn on K-12 summative high-stakes tests. We discuss the implications given different perspectives of the role of testing in education.

Introduction

The United States Office of Educational Technology (2017) issued a national technology plan updatethat included an emphasis on personalized learning. The report states:

> "Personalized learning refers to instruction in which the pace of learning and the instructional approach are
> optimized for the needs of each learner. Learning objectives, instructional approaches, and instructional content (and its sequencing) may all vary based on learner needs. In addition, learning activities are meaningful and relevant to learners, driven by their interests, and often self-initiated." (p. 9).

In alignment with this definition of personalized learning, the report emphasizes the importance of embedded assessments in educational technology platforms (United States Office of Educational Technology, 2017). Embedded assessments are essential in educational technology platforms if one is to personalize education, as assessment is critical to track the learning of the student. However, the report also briefly discusses the importance of summative statewide standardized tests in U.S. public schools for achieving accountability and equity in schooling.

The notion that schools can fully adopt personalized learning methods while operating under current federal summative test accountability policies is questionable. Hyslop and Mead (2015) detail this tension. Personalized learning allows for differential student paths and paces through a curriculum, whereas U.S. accountability systems monitor student learning in alignment with grade-level expectations that are consistent across all students. These expectations are assessed at a single time point (i.e., end of year), based on the premise that all students should meet grade-level learning standards by that time point. Yet the theory of personalized learning aligns with assessments that monitor growth over time within students in tandem with their individualized pace of learning, rather than comparing all student learning to the same grade-level expectation at the same time.

Over the past five years, our research team has been implementing personalization algorithms based on artificial intelligence (AI) in a virtual learning environment for Algebra instruction with high-school students in the state of Florida (Leite et al., 2022). A primary goal of the research is to gauge if the personalization is effective for preparing students to pass the high-stakes, Algebra 1 End-of-Course (EOC) assessment in the state. Students must pass this exam to graduate (Florida Department of Education, 2021). We entered the research project theorizing that a more personalized approach to Algebra learning and test preparation may increase the percentage of students passing this high-stakes exam. However, we quickly came to see the tensions between personalized learning and state accountability and testing systems discussed by Hyslop and Mead (2015). How can we ensure that all students are ready for the same algebra test at the same time, while allowing a high degree of freedom in the personalization of algebra learning? What are the consequences for students if they enter the high-stakes test at the same calendar time, but at a different point in their personalized paths through the curriculum? Is it fair to submit all students to the same high-stakes test at the same time after we have implemented personalization algorithms that have likely introduced content exposure and opportunity to learn variability amongst students within particular curriculum units? From a measurement perspective, should we conceptually make claims about student math abilities based on an end of year exam when those students have had different paths through the curriculum? These concerns motivated us to conduct this study.

The purpose of this study is to empirically evaluate the differences in Algebra 1 content exposure at the time of the Florida Algebra 1 EOC assessment administration across groups of students who vary in personalization paths within the curriculum. The *Standards for Educational and Psychological Testing* (2014) are clear that unequal opportunities to learn test material prior to taking tests can be a problem for fair measurement. Hence, a large variance in Algebra content

exposure across students at the time of the high-stakes assessment poses concerns for fairness in interpreting and using the test scores as a major decision point for earning a high-school diploma. As our larger study is ongoing and has experienced issues with both fidelity to personalized recommendations and temporary changes to high-stakes assessments during the 2020–2022 COVID-19 health pandemic (Leite et al., 2022), we focus our study on a hypothetical, simulated situation in which students in Florida engage with our personalization algorithms exactly as intended in the algorithms' construction. The overarching research question of our simulation is: After engaging in the AI-enhanced curriculum for a full school year, what are the Algebra 1 content exposure differences amongst students who have received personalized instruction as well as students who have received non-personalized instruction?

The significance of answering this research question is three-fold. First, while researchers and practitioners have discussed the tensions between personalized learning and U.S. accountability policies (e.g., Hyslop & Mead, (2015)), we have not seen empirical studies that inform the degree of this tension. In conducting our study, we aim to quantitatively evaluate the degree to which differential content exposure is expected when personalized learning is enacted in schools. Second, issues around unequal content coverage prior to taking a high-stakes test have historically been handled in the legal system, and the evidence of content coverage has stemmed from teacher descriptions of the course curriculum in general, which are often self-reported and not specific to any one student (e.g., *Debra P. vs Turlington*, as discussed in Pullin & Haertel, 2008). Our research project provides an example of a benefit of using AI-enhanced curricula, in that we can more accurately track and evaluate content exposure for each student prior to taking high-stakes exams. We hope our study serves as an initial example for more formally evaluating opportunities to learn through content exposure for all students and provides a methodology for considering such

questions in the future that may involve different personalized learning algorithms. Third, the health pandemic and societal racial reckoning that upended U.S. society in 2020 called into question the fairness and usefulness of large-scale standardized testing in general (see for instance, Jiao & Lissitz, 2020). Hence, we see this study as contributing to the broader reexamination of equity and fairness in educational measurement systems.

An overarching aim of this article is to highlight the different perspectives one can take about what constitutes test fairness. For instance, in the realm of dynamic assessment, the ideal for test fairness is one that is individualized and learner centric. Baharloo (2013), for instance, notes that most common views on test fairness, whether related to measurement validity (for instance, Kane, 2010) or even from a social justice perspective, are often focused on more traditional assessments in which learners and test takers are expected to have common knowledge and common preparation. However, the forms of test fairness from which dynamic assessment is derived says that a test or assessment should blend teaching, learning, and testing to help learners develop in their own way (Baharloo, 2013; Poehner, 2011). The goal of testing in dynamic assessment is to make teaching and assessment part of the same task, and fair assessment is viewed "in terms of providing each individual with what he or she deserves based on need analysis and ongoing assessments used for diagnostic purposes" (Baharloo, 2013, p. 1936). In this way, the two views of fairness can be seen in potential opposition and promoting two different forms of testing that are value-laden and specific to the purpose of the test. What is considered unfair in one realm may be considered fair in the other realm of testing.

The conflicting two views on testing mentioned in the paragraph above are perhaps best represented by the chapter on fairness in the *Standards*, which emphasizes equal opportunities to learn, and the 2017 United States Office of Educational Technology statement about personalized

learning that opened this paper. More broadly, from a measurement perspective, these differences may additionally manifest in test score interpretation. For instance, one may worry about the extent to which content exposure differences should be considered part of the property to be measured or something external to the property that is to be measured or assessed.

1.1. Virtual Learning Environments, Personalized Learning, and Test Fairness

With most students in the United States now having access to computers and the internet at home, the use of online learning systems has been increasing (see for example, Koedinger et al. 1997 for an early discussion of online learning systems, Mean et al., 2013, for discussion of their spread and use from a policy perspective). There are many types of online learning resources for students, with some being designed to track student progress through online content that is specifically built around a theory of learning as well as a specific curriculum in the K-12 system. In some cases, these systems are designed for test preparation. The online learning platform of interest in this paper is one such case, and we refer to these types of systems as virtual learning environments (VLEs; Weller, 2007).

Some VLEs are advanced learning technologies (Aleven, Beal, & Graesser, 2013) that use AI to personalize the students' learning experience by matching specific learning resources to predicted student needs. When personalizing VLEs to individual students using AI, it is critical to understand if and when the students will be required to take a test on material that has high-stakes for the student. The *Standards for Educational and Psychological Testing* state:

> When student mastery of a delivered curriculum is tested for purposes of informing decisions about individual students, such as promotion or graduation, the framework

> elaborating content domain is appropriately limited to what students have had an opportunity to learn from the curriculum as delivered (*Standards*, 2014, p. 15).

This formal standard for testing makes one thing clear: if a student has not had basic exposure to all curriculum units covered by a high-stakes test, the validity of the high-stakes decision is questionable. Extending this idea, if some students have had more opportunity to learn from particular units of curriculum content than other students, yet all are taking the same high-stakes test with the same high-stakes decisions criteria, then test fairness is questionable.

There has been a long history of trying to understand and study opportunity to learn. For instance, Hume and Coll (2010), Kurz, Elliott, Kettler, and Yel (2014), and Kurz, Talapatra, and Roach (2012) note the effects of opportunity to learn on student achievement. Kurz et al. (2014) discuss that Carrol (1963) was the first to provide some sort of workable definition, or framework, for considering opportunity to learn. In this simple definition, opportunity to learn is considered from the perspective of time devoted to topics within a curriculum as well as time devoted to content. However, one can see that the frameworks considering quality of coverage (Kurz et al., 2014) for opportunity to learn are also important (how might one learn, say, if content is covered but too quickly or the environment for learning is poor?). Using these concepts, there are a variety of aspects that one may consider important for study – though, it is likely that not all can be considered. Kurz et al. (2014), for instance, tried to create an index of opportunity to learn using online teacher logs that covered time, engagement, and whether teachers covered certain cognitive processes or something similar. However, in the context of VLEs in which personalized learning systems are used, there are other considerations – such as how a system makes recommendations based on the personalization parameters. Additionally, understanding how students move through a VLE can be informative of what students may find

useful, or what is reasonable to expect from students given a current online learning system setup – for instance, their time spent on consistent work (see, for example, Lowes, Lin, & Kinghorn, 2015).

As stated, the proliferation of VLEs with some element of personalized learning places students on different learning trajectories. Students may see different content by the time of the high-stakes test. In the context of this study, the Algebra 1 EOC exam is administered at a single time point for a high-stakes decision about each student. Therefore, there are numerous questions about fairness and validity of decisions made from scores on the EOC when personalized learning is the primary mode of content exposure and learning. From a content exposure perspective of opportunity to learn, it is not clear that this system is fair. Therefore, crafting both conceptual and analytical strategies for understanding opportunity to learn is important – especially as there is an increasing number of VLEs used in grades 3 through 12 classrooms that are subject to United States federal accountability testing policies.

1.2. Algebra Nation and the Personalized Recommendation System

Algebra Nation (AN), which is embedded in Math Nation (Lastinger Center for Learning, & University of Florida, 2019), is a VLE designed to facilitate the instruction of algebra and to prepare students for passing a state-wide Algebra 1 EOC exam required by the state of Florida for graduating high school. Every student and every teacher in Florida have access to AN using their district-assigned user id and password. There is correlational and quasi-experimental evidence that student and teacher use of AN is associated with increased scores on Algebra 1 EOC (Leite, Cetin-Berber, Huggins-Manley, Collier, & Beal, 2019; Leite, Jing, Kuang, Kim, & Huggins-Manley, 2021; Niaki, George, Michailidis, & Beal, 2019).

AN consists of instructional videos with 10 sections covering different algebra content. There are multiple topics and topic videos within each section, for a total of 93 topics in AN. In our simulation, due to problems related to items for assessing two topics, we constrained our simulation to 91 complete topics. For each topic video, students can select among different tutors of different ethnicities. These tutors use intentionally different presentation styles, where some tutors are more thorough – for students who are newer to a topic or need more instruction - and some tutors are faster, treating their videos as review sessions. The videos vary in length, being anywhere from 5 to 30 minutes. After watching a video on a topic, students take a "Check Your Understanding" quiz (CYU) consisting of three items. These items are built to be similar to those one might find on the EOC assessment.

The current operational version of AN does not include personalization. We implemented AI-based personalization for students in a single school district as part of a larger study (Leite et al., 2022). The CYU items used to guide personalization in AN have been field-tested and pre-calibrated (see Xue, Huggins-Manley, & Leite, 2021, for details) under a two-parameter logistic item response theory model (2PL; Birnbaum, 1968). The system calibrates the student trait score on the CYU test using *expected a posteriori* (EAP) estimation with fixed item parameters, and then compares the trait score to an average trait score on that CYU test from other students in the system that we call "peers." Peers were determined by stratifying the achievement data based on Mahalanobis distances between the students prior to entering the personalization. The recommendation for the next action is based, in large part, on the distance from the student's EAP estimate on the CYU to other students with similar academic backgrounds on that same CYU. Students also receive an engagement score based on time spent and actions in the platform (see Jensen, Hutt, & D'Mello, 2019 for details on quantifying engagement). The CYU score in

relation to peers and the engagement score both play a role in the recommendation, along with video topic weights. Each video topic in the system is associated with a time-invariant topic weight that represents the relative importance of given topics to passing the EOC based on previous data (Leite et al., 2022). Using all of this information, a probabilistic model is used to make a video recommendation with certain codes and rules. However, this recommender is only used in cases where students get 0 items or 1 item correct on the CYU (out of 3). In these cases, the recommendation system may recommend a new video, rewatching a video, rewatching a portion of a video, or rewatching a video with a new teacher as the speaker. The recommendation can be forward or backwards in the progression of the videos and algebra topic. If students get all CYU items correct, they are directed to move linearly to the next video. If students get two items correct, students are instructed to re-watch the video. One can see that given a set amount of time, students who follow recommendations may see fewer topics because they can repeat topics and videos. However, the reinforcement learning algorithm used, based on a theory of learning, is meant to help with mastery of topics. More details on the full recommendation system are provided in (Leite et al., 2022).

The version of the AN platform we studied is a tailored, active VLE motivated by the notion of personalized learning for students. As a result, AN is expected to lead to differential content exposure among students with the idea that students will see more of the content they need to see to pass the exam. However, all students are required to take the Florida Algebra EOC at the same time point, so students may not have been exposed to and mastered all content. Hence, this study uses a simulation of this system (in which all students have the same amount of time within the VLE) to answer the overarching research question: After engaging in the AI-enhanced curriculum for a full school year, what are the Algebra 1 content exposure differences amongst

students who have received personalized instruction as well as students who have received non-personalized instruction? The specific sub-research questions under this are:

1. What are Algebra 1 content exposure differences between students who receive personalized instruction and those who do not?
2. Among these groups, how does student learning throughout the system affect student content exposure?
3. How do different student math abilities relate to students' content exposure?
4. What do the paths through the content in the system look like for different students?

2. Method

We conducted a simulation study that was intended to mimic the actual operation of the personalization system as designed and presented above. We created and compared three groups of hypothetical students.

1. Personalized growth group: Students who received topic and video recommendations and whose CYU trait scores grew within each section each time they say a new topic.

2. Personalized non-growth group: Students who received topic and video recommendations but did not grow in CYU trait scores within each section.

3. Control group: Students who did not receive topic or video recommendations but simply advanced through the system linearly, moving from one topic to the next in order of their presentation in AN, which aligns with the state algebra standards progression.

Assumptions had to be made to commence the simulation, and we list them out here as they influence the answers to our research question.

Assumption 1: All students complied with recommendations.

Assumption 2: All students would spend two hours within each section. That is, the simulation had students work through videos and take quizzes for a total of two hours within a section before proceeding to the next section.

Assumption 3: Students selected among the video teachers probabilistically based on empirical data. That is, based on observed student selection of tutors in real data, the probability of a student selecting a particular tutor was equal to the proportion of students who selected the tutor in real data.

Assumption 4: Students would take approximately 2 minutes per quiz question. The implication is that the three-item item quizzes that students took at the end of each topic took 6 minutes.

Assumption 5: Student engagement was constant within a section of AN.

Assumption 6: The topics increase in difficulty ordinally. That is, topic 1 is "easier" than topic 2, and topic 2 is in turn "easier" than topic 3.

Assumption 7: Simulated students from the growth condition grew .1 logits in CYU trait scores each time they saw a new topic but returned to their starting CYU trait score when they entered a new section.

Assumption 8: Control group simulated students would always move one topic video forward and never backwards or skipping a topic, regardless of their CYU trait scores.

2.1. Generating Students for the Simulation

For simulating hypothetical students, we began by defining 20 clusters of students that represented peer trait score groups. These clusters were generated such that cluster 1 had the lowest average trait score estimate and cluster 20 had the highest average student trait score estimate. In each cluster, student abilities were generated from normal distributions with the mean and standard deviation equal to the cluster mean and standard deviation. These clustered "peer" students were used to create average abilities for each topic within each section. See Table 1 for an example of resultant trait score estimates. These cluster trait score parameters (in logits) mimicked what was found in operational data.

Table 1

An example of the peer trait score clusters where values represent average-estimated trait score for that cluster for each topic in Algebra Nation (in logits). You can see that cluster one has the lowest values and average trait score values ascend from there.

Topic #	clust_1	clust_2	clus_n	clust_19	clust_20
1	−1.24	−1.14	.	0.71	0.68
2	−0.52	−0.60	.	0.43	0.68
3	−0.93	−0.79	.	0.49	0.40
4	−1.23	−1.10	.	0.65	0.55
5	−0.93	−0.68	.	0.59	0.56
6	−0.87	−0.87	.	1.13	0.97
7	−0.93	−0.91	.	0.77	0.78
8	−0.77	−0.73	.	0.83	0.79
.	.	.	.	.	.
.	.	.	.	.	.
89	−0.68	−0.73	.	0.45	0.48
90	−0.58	−0.45	.	0.46	0.70
91	−0.39	−0.39	.	0.29	0.51

From each cluster, we drew 100 students randomly, generated from a normal distribution with the specific cluster parameters describing the distribution from which to be drawn. The trait score drawn for the student was subsequently treated as the true trait score of the student, with the exception that in the "growth group" this true trait score was increased by .1 logits within a given section for each new topic that was presented. This growth occurred across topics within each section, but when entering a new section, the original true trait score value was used again (without the growth) and growth occurred again within that new section. Each student was then

individually sent through the recommender system. We acknowledge that the number of assumptions here is large – however, we believe this method is useful for modeling the effects of different assumptions that can eventually be investigated empirically and used to update the simulation.

2.2. Generating Item Response Data and Estimating Student Abilities

Since the recommender system only estimates a student's trait score level based on three items, we generated item responses on the pretest for each section based on individual student true abilities. To generate this data, we used individual student abilities along with the already pre-calibrated item difficulties and discrimination parameters for the AN items, and generated response probabilities for an item. The probabilities for each student and item were compared to a random number drawn from a uniform distribution that ranged from 0 to 1. If the response probability was less than the drawn number, the individual response was coded as incorrect, or 0, and 1 otherwise. Then, we estimated student abilities from the generated item response data from the CYU quizzes just as would be done in the AN platform with real students providing correct and incorrect responses on the items. This was recorded as a student CYU trait score for that topic. This full process was conducted whenever a simulated student was exposed to a new topic under the personalized recommender system. For control group students, no trait scores were needed as they do not play a role in their path through the system.

2.3. Length of the Simulation

Video times from within AN were recorded and used for setting how long a student was in the system. As noted above, a student was estimated to be in the system for two hours. We assumed students took two minutes per test item. We also generated engagement scores randomly to mimic the distribution of engagement scores in operational data. When a student had been in the system for two hours, students moved onto the next section. However, students could be recommended a topic so that the student is moved backwards or forward across sections. The operational recommender system constrains how far forward or backwards a student could move based on how far along the student is in the school year and what other videos the student has already viewed. To mimic this, since AN serves as supplement to ongoing instruction and corresponds to a student's position in the school year, AN sections were effectively equated to progression through the school year (so, the first section, students were $1/10^{th}$ of the way through the school year, second section, $2/10^{th}$, etc).

2.4. Evaluation Criteria

For answering our research questions, we recorded the video and topic covered. From a coverage perspective of opportunity to learn (as opposed to mastery) this codifies if a student had the opportunity to learn from a video or not. This was simply coded as 0 or 1. See Table 2 for an example student's results matrix. In the results matrix in Table 2, each topic corresponds to a topic in AN. If a student was recommended and watched a topic, the "seen" column was coded as 1. This does not record how many times a student saw a topic and it does not record in what order the student saw topics. Additionally, we recorded the raw topic data. This raw topic data kept the order of the topics viewed and included repetition. For instance, if a student first saw

topic 3, and their next recommendation was topic 3, then both instances were recorded. In this way, we could chart student paths through the system given the simulation parameters.

Table 2.

An example student results matrix that was analyzed for assessing opportunity to learn

topic	seen
1	1
2	0
3	1
.	0
.	1
.	1
.	0
89	1
90	1
91	1

3. Analysis

Research question 1 and 2 were answered using two methods. The first used descriptive statistics. We estimated mean content coverage across the three primary simulation conditions and visualized the content exposure differences across different learning and recommendation conditions. For the descriptive inferences, we estimated the proportion of the EOC exam content a student was exposed to in AN by using a weighted sum of the topics to which each student in the simulation was exposed. That is, each topic in AN received a weight, as mentioned in the paper introduction, based on the given topic's prevalence on the EOC exam. The sum of topic weights added up to the total number of topics so certain topics had weights greater or less than one. These topic weights

were multiplied by 1 if a student saw a video about a topic and a zero otherwise, summed, and then divided by the total number of topics used in the simulation (91). This calculation represented the proportion of the EOC exam to which a student was exposed in AN. For the visualizations, we plotted the proportion of each section of AN that a student was exposed to. Exposure was defined as the number of videos a student was recommended within each section divided by the total number of videos in the section. In the second method, which will also help answer research question 3, we used a repeated measures logistic regression approach. To answer our questions, we regressed the outcome of interest, whether a student covered a topic or not, on six predictors (variable name in parentheses):

1. AN section (one through ten) (section)
2. Topic number within section (topic)
3. Engagement value for a student measured at topic *t* (integer value from 1 to 5; engage)
4. Peer cluster number (1 to 20; cluster)
5. CYU trait score relative peer group trait score, standardized (rel_abil)
6. Learning condition (personalized growth group, personalized non-growth group, control group; learningCondition).

The full model is defined as:

$$logit[P(Covered_{st} = 1)] = \beta_1 section_{st} + B_2 topic_{st} + B_3 engage_{st} + B_4 cluster_s + B_5 rel\_abil_{st}$$

4. Results

4.1. Differential Topic Exposure

To answer research question 1, we first estimated the proportion of EOC content to which the student was exposed. On average, students in the control group covered 61% of the content that would appear on the EOC. Alternatively, both the growth and no-growth groups (both groups receiving recommendations) were exposed to an equal 46% of the content that would appear on the EOC.

For the visualizations, we present box plots intended to convey the difference in content exposure between three student groups (not weighted for importance on the EOC). Figure 1 shows the section number and the corresponding subject covered within that section on the x-axis and the proportion of the total number of videos covered by students. One notes, that almost always, the median proportion of a section covered from the control group – the group that moved sequentially with no recommendation system – was higher than the recommendation system students. The range and variance of the proportion covered are also worth noting. In some cases, the control group has only a few combinations of data for proportion covered (see sections 5, 7, 9, and 10). These are cases where differences in video times between speakers is relatively low. In the other sections, we can see that, aside from section 6, the range of the control group is often much smaller than the recommendation groups (ignoring "outliers" – which in the control group are caused by speakers who have much longer videos in specific topics). However, for the recommender groups, the minimum and maximum content exposure in some sections was 0% and 100% of videos viewed, respectively – since the recommender can

recommend videos within and outside the current section to maximize student mastery.

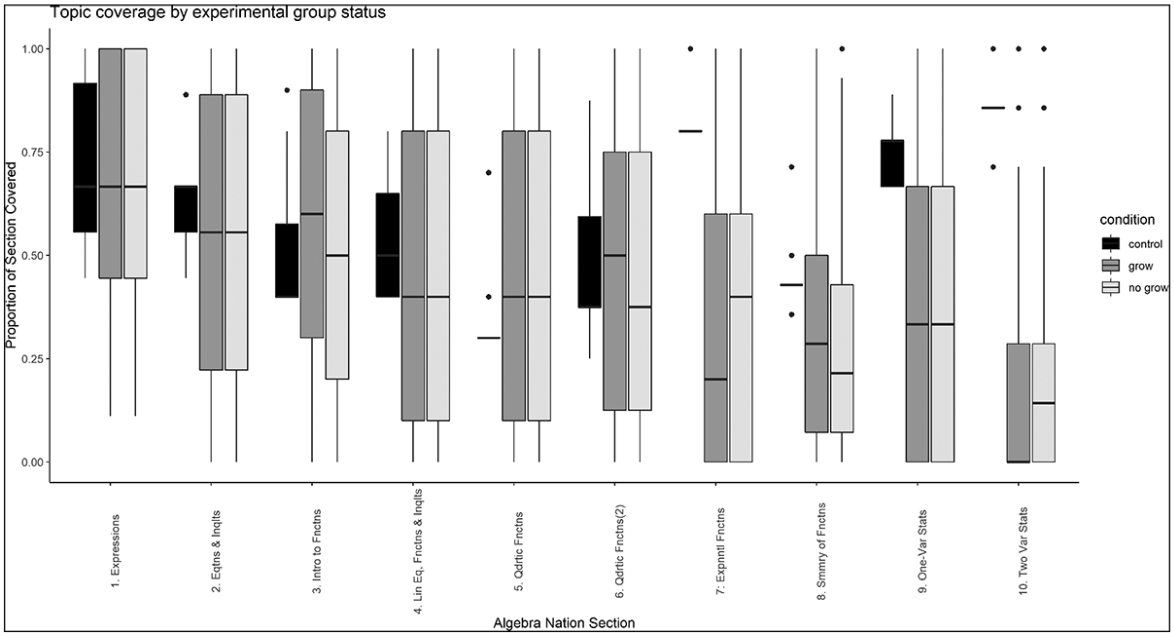

Figure 1 Distribution of content exposure across learning (grow) and no learning (no_grow) groups as well as the control group.

Since we are additionally interested in the effect of student peer cluster on content exposure, Figure 2a-2d shows the differences content exposure for four clusters individually and the experimental conditions within each cluster. The figure visualizes the lowest trait score group (cluster 1), the middle trait score group (cluster 10), and two higher trait score groups (clusters 15 and 20). One can see that while the inferences about control vs. recommendation (growth and no growth) simulation conditions hold within cluster as well as overall (see Figure 1), we can also describe specific features related to the highest cluster of students (i.e., those with higher CYU trait scores throughout the platform experience). It seems like students in the highest

cluster are more likely to be sent backwards since slightly more students, based on the 25th percentile line in the box plots, are likely not to have exposure to certain sections starting with section 5. For instance, Figure 2a, showing the proportion of sections covered for cluster 1, only sections 9 and 10 (the last two on the x-axis) have a significant number of students seeing no portion of those sections. However as the cluster number increases, and hence student trait scores, a greater number of sections have 25% of students not seeing those sections. Figure 2d, for cluster 20, the highest ability group, has five sections that show this pattern (sections, 5, 6, 7, 9, and 10). This helps reveal the logic of mastery within the personalization system – moving forward, or at least seeing videos associated with later (and more advanced) topics, is not necessarily viewed as beneficial for high trait score students. In other words, cluster coverage does not seem to increase across sections as trait score increases. This is descriptive evidence that routes though curriculum can be qualitatively different (ideally informed by student learning paths). Having said that, the logisitic regression results reveal that the impact of clusters on EOC content exposure is not large.

The logistic regression model, addresses research questions 1 and 2 but especially targets research question 3. Table 3 reports the results of this model and these results are discussed further below.

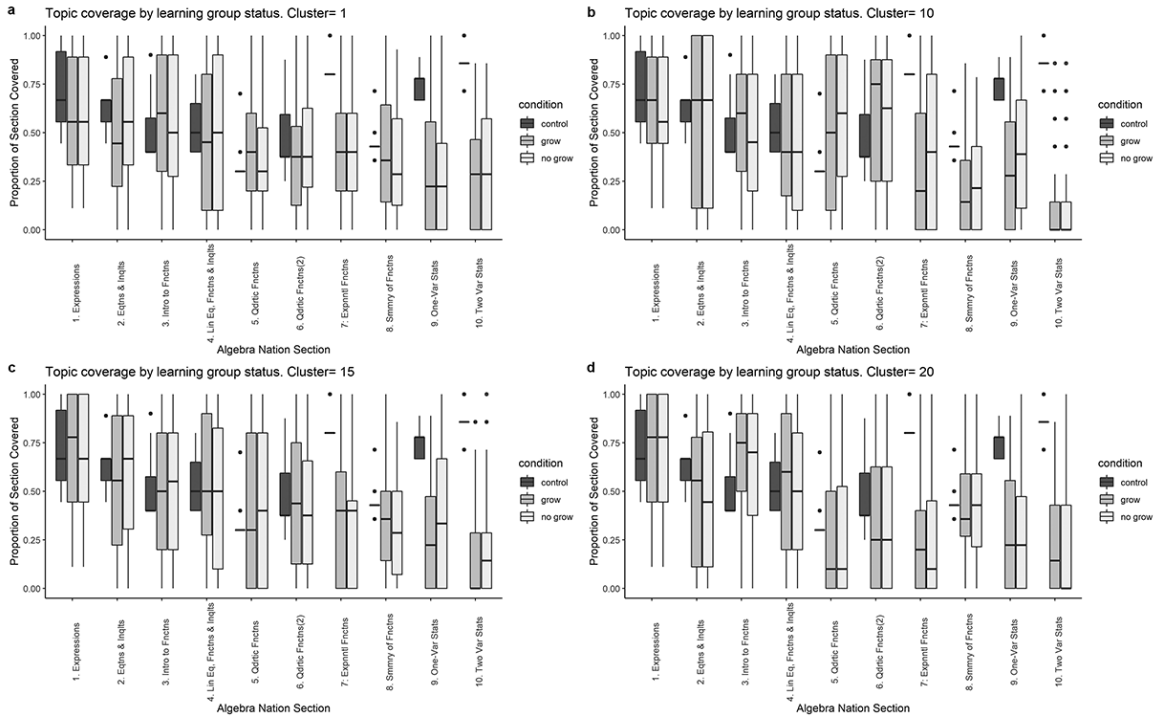

Figure 2 The distribution of content exposure in Algebra Nation across clusters and growth conditions. The top panels (a) and (b) show students across simulation conditions of the lowest trait score and middle trait score and the lower panels(c) and (d) show students across conditions in the top trait score clusters (15 and 20), shows students across conditions in the highest trait score cluster.

Table 3

Logistic regression results with parameter estimates presented as odds ratios with the personalized growth group as the reference group

	Estimate (OR)*	SE	Est/SE**	p_val
SECTION	0.838	0.002	−97.409	< 0.001
TOPIC_WITHIN	0.936	0.001	−45.315	< 0.001
ENGAGE	0.975	0.005	−4.734	< 0.001
CLUSTER	1.002	0.001	3.779	< 0.001
REL_ABIL	1.007	0.003	2.221	0.026
NO GROW	0.992	0.006	−1.332	0.183
CONTROL	1.969	0.041	23.457	< 0.001

* OR = odds ratio **Est/SE =(1 - exp(Estimate))/SE

4.2. RQ 1: Topic Exposure – Personalized Vs Control Groups

The control condition students were approximately two times more likely to cover any given topic than either the personalized growth or no growth conditions of students, adjusting for other predictors in the model (odds ratio = 1.97, $p < .001$).

4.3. RQ2: Learning Conditions – Personalized Growth Vs Personalized No Growth Groups

The learning condition (given our assumptions) did not significantly affect topic exposure. Students in personalized groups were no more likely to see a given topic if they were learning than not (after adjusting for other covariates). This finding may not hold if student learning was larger than the 0.1 logits per topic that we used for this simulation. i.e., .1 logit increase with each topic increase).

4.4. RQ3: Math Abilities and Content Exposure

Though each additional cluster number (such that an increase in cluster number represents a higher overall cluster trait score) is associated with an increase in the odds of covering a harder (later in the section) topic (OR = 1.002, $p < .001$), this associated increase is extremely small. So, while it may be the case that a student in cluster 20 is more likely to cover a topic than a student in cluster 19 the differences in paths through AN might be more interesting in kind than in quantity. Students' relative trait score within cluster also had a small impact on content coverage. For example, for a 1 SD difference in trait score relative to the cluster mean and standard deviation, the odds of students seeing a topic were only nominally larger (less than 1% more likely, OR = 1.007, $p = .026$).

4.5. RQ4: Student Path through AN

Another relevant aspect for teachers who are considering instructional strategies combined with the use of an online learning system is student paths through the system. For this study, we are hoping to present methods for thinking about or presenting this information. Further, it is worth

considering a more complex view of opportunity to learn that combines not just whether a student viewed material, but when and in what contexts. To present a few examples of these plots that support the general theme that often student exposure in AN is different not just in quantity but also in profile, alluvial plots are presented in Table 3. The data in these plots are filtered such that only topics with more than 20 views across all clusters are presented (otherwise, the alluvial plots become unreadable). Additionally, these plots show only the personalization groups, and notably aggregate the data across the growth and no-growth conditions.

These paths are presented in Figure 3 for students in the three different clusters, namely the same clusters as presented above in Figure 2. Each plot shows the students' paths for Section 5 only. One can see that while there is topic overlap, in a single 2-hour session, students in different clusters or even within clusters may have completely different paths through the system. So, in this sense, their opportunities to learn through content exposure are qualitatively different. For instance, for the lowest trait score cluster, one can see more homogeneity in their paths through the system - denoted by the "alluvials" or thicker blocks. Thicker blocks and alluvials generally represent more simulated students being recommended a particular video topic - for instance, cluster 1 students in section 5 were primarily presented topic 41. Their most common second topic to see, denoted by the thick flow to the next topic, is to watch topic 41 again. However, of note, students in cluster 20 have much more varied topic combinations through AN's section 5 than students in cluster 1 or cluster 15 (especially for the first two moves) implying that if a test were to be given at the same time point across these students, we can expect a large degree of differential opportunity to learn through content exposure.

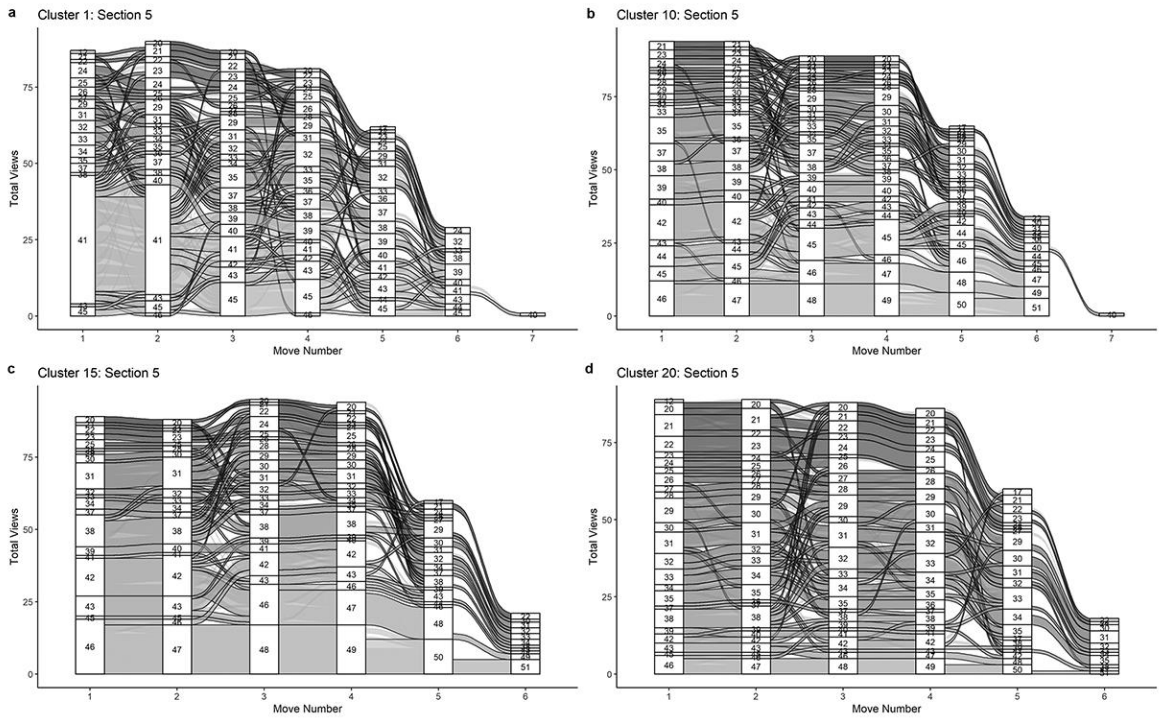

Figure 3 Alluvial plots representing common paths through Algebra Nation's Section 5, where the x-axis represents each move in the system a student would make and stacked numbers represent topics. The thicker "alluvials" and flows, represent more common paths for each cluster. Panel A depicts paths through Algebra Nation in Section 5 for cluster 1, the lowest trait score panel. Panel B depicts cluster 10 paths, Panel C represents cluster 15 paths, and panel D presents cluster 20 paths.

5. Discussion

In the current study, we presented hypothetical results from different fully complying students moving through an online learning system that tailors content to a student's learning. In order to make recommendations to students, the AN system (Leite et al., 2022) uses embedded quizzes to estimate student trait scores for that topic and also estimates a student's engagement. Based on a

theory of learning and AI algorithms, a topic recommendation is made to the student. This study began by noting an inherent problem – while standard testing practices and guidelines, as those noted by the *Standards*, require that tests for decisions about selection at a given point in time require that students have an equal opportunity to learn, the growing number of personalized online learning platforms lead to students having potentially different curricular exposure by the time of the test. Therefore, this study highlights a disconnect between standard guidelines for testing and new educational technology that takes advantage of data mining and machine learning methods (e.g., Baker & Yacef, 2009; Koedinger, D'Mello, McLaughlin, Pardos, & Rose, 2015).

While we invoked many assumptions for the purpose of this simulation, we tried to be as transparent as possible about these assumptions. The reason for this transparency was both for reporting purposes and also for identifying which assumptions to investigate in the future. As expected, students who received recommendations saw fewer topics than those who moved through the system linearly. However, students who entered the system with higher algebra abilities were actually more likely to be sent "backwards" in the system, and, as can be seen descriptively, may have more variance in topic exposure than those students who did not receive recommendations or were of lower math ability. Additionally, while it is likely that students in higher ability peer groups see more topic videos, the more interesting differences in topic exposure may be differences in kind than in quantity. However, student growth in ability seemed to have only a small effect on topic exposure, if any. The student paths through the system may be not only more interesting, but also more useful for teachers in knowing which students saw specific topics. So, in this sense, their opportunities to learn are qualitatively different.

The primary goal of this study is not *just* to show what would happen to students if they moved through the system, but to also present a method for interrogating a probabilistic recommendation

system that is based on measurements occurring within the system. As noted by Mislevy (for instance, Mislevy et al., 2013), assessments often need to be viewed as part of a system – and the system of test preparation is considered important and relevant. This preparation may be viewed as a threat to test validity or a feature of test validity depending on how the test is built. In the case of AN, the system was specifically built to prepare students for an exam. However, it is at odds with usual perceived beliefs about students. It is clear, then, that even tacitly, summative end of year tests are part of a learning system and do not necessarily stand as isolated events. As such, this paper is meant to demonstrate that measurement professionals always need to consider this system holistically.

We additionally hope that this paper serves as a starting point for any other researchers looking to understand the extent of a recommendation system's impact on fairness and equity of assessment. Being able to understand the implications of different recommendation systems on measurement fairness, especially for downstream measurement efforts, will likely become increasingly important as online learning systems become more common. Measurement fairness in AI-based VLE should become part of the broader conversation about algorithmic fairness in education (Kizilcec & Lee, Forthcoming; Madaio, Blodgett, Mayfield, & Dixon-Román, 2021). Hopefully simulations like this one can serve to motivate policies around testing and the role of testing. Dorans (2012) notes that there are multiple perspectives one can take about the same test that may differ among test takers and other stakeholders. This perspective taking may change whether the student or stakeholder views the test as a contest of some sort or as a measurement. The way a test like an end of course exam is administered and its role in student learning may also change depending on the system leading up to the test. *In our study, the system is meant to promote and maximize current learning whereas the end of course exam is not meant to do the same*. It is

also yet possible that, combined with classroom instruction, an online learning system could increase exposure to certain topics deemed important for passing tests.

Finally, this study provides a starting point for understanding how to maximize student use of the VLE. First, the simulation can be used in the future to understand how different assumptions may change the recommendation. Also, it may allow for presenting hypothetical results to teachers to better understand how to change the system to better align with classroom-based practice, within a broad strategy of orchestration of technology in the classroom (Prieto, Dlab, Gutiérrez, Abdulwahed, & Balid, 2011). Since student use is widely varied and voluntary, having an ideal scenario to compare to actual use provides opportunities to make updates to the recommender, re-simulate, and compare again to actual student use – each time considering the curricular, test, and measurement effects. Lastly, we hope to motivate others to use this simulation set up to assess the fairness of their online learning system with respect to measurement.

Moving forward, we hope to both interrogate some of the necessary assumptions based on empirical data and use the simulation results to inform future work. While there are certainly many potential technical answers to the optimal design of an online learning platform, we hope that this discussion also reveals the value laden aspects of educational measurement – as these values are embedded in the way test preparation may occur. The values of this test preparation proliferate or spread to the test.

From the perspective of fair testing writ large, this work contrasts different ideals in the realm of educational measurement, such that what may be unfair in one world may be fair in another, simply based on value-based decisions of which measurement forms one aspect (e.g. Lantolf & Poehner, 2013). For instance, Nisbet and Shaw (2019) highlight different perspectives on test fairness. Among the perspectives, they emphasize that such standard tools in educational

assessment like differential item functioning (DIF) analyses are used to test hypotheses about a view of fairness in which fairness is established by ensuring the comparison of like cases. However, Nisbet and Shaw (2019) note that this is only one of many perspectives on fairness we could take. In personalized learning, cases are by design different, raising the question about what constitutes the primary student property to be assessed or measured. For instance, from an end of course exam perspective, differential content exposure would lead to construct irrelevant variance (Messick, 1989) – a different causal mechanism would lead to differing test scores by student, confounding the property of interest, "math ability" perhaps, with background knowledge. In personalized learning, a different causal pathway is something to be desired and engineered in some cases if it aids the student in their learning progression. The perspective of Randall (2021) provides another important perspective on the view that different causal pathways to test scores should be considered from an anti-racist perspective - and not doing so is another source of unfairness.

Given the rise of personalized online learning while continuing to operate under end-of-year accountability testing policies, fairness of measurement claims is a topic with which educational measurement researchers and professionals must grapple. Differential content exposure complicates already notable paradoxes in testing related to measurement invariance and prediction invariance that should be explored further (see, for instance, Borsboom, Mellenbergh, & Van Heerden, 2002 or Zwick, 2019). It is hoped that the results from this study help show the implications of these conflicting views.

References

Aleven, V., Beal, C. R., & Graesser, A. C. (2013). Introduction to the special issue on advanced learning technologies. *Journal of Educational Psychology*, 105(4), 929–931.

Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies*, 3(10), 10.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A Review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick's (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, Mass: Addison-Wesley.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, 26(4), 433–450.

Carrol, John B. A model of school learning. *Teachers College Record*, 64, 723- 733

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, 31(4), 20–37.

Florida Department of Education. (2021). Florida standards assessment. http://www.fldoe.org/accountability/assessments/k-12-student-assessment/fsa.stml

Hume, A., & Coll, R. (2010). Authentic student inquiry: The mismatch between the intended curriculum and the student-experienced curriculum. *Research in Science and Technological Education*, 28(1), 43–62.

Hyslop, A., and Mead , S. (2015). *A path to the future: Creating accountability for personalized learning.* Bellwether Education Partners.

Jensen, E., Hutt, S., & D'Mello, S. K. (2019). Generalizability of Sensor-free affect detection models in a longitudinal dataset of tens of thousands of students. In M. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou (Eds.), The 12th International Conference on Educational Data Mining, Montreal, Canada (pp. 324–329).

Jiao, H., & Lissitz, R. W. (2020). What hath the coronavirus brought to assessment? Unprecedented challenges in educational assessment in 2020 and years to come. *Educational Measurement: Issues and Practice,* 39(3), 45–48.

Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182.

Kizilcec, R. F., & Lee, H. (Forthcoming). Algorithmic fairness in education. In W. Holmes, and K. Porayska-Pomsta (Eds.), *The Ethics of Artifical Intelligence in Education*. Taylor & Francis.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rose, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Review Cognitive Sciences*, 6(4), 333–353.

Kurz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment*, 19(3), 159–184.

Kurz, A., Talapatra, D., & Roach, A. T. (2012). Meeting the curricular challenges of inclusive assessment: The role of alignment, opportunity to learn, and student engagement. *International Journal of Disability, Development and Education*, 59(1), 37–52.

Lantolf, J. P., & Poehner, M. E. (2013). The unfairness of equal treatment: Objectivity in L2 testing and dynamic assessment. *Educational Research and Evaluation*, 19(2–3), 141–157.

Lastinger Center for Learning, & University of Florida. (2019). Algebra nation. Retrieved September 20, 2019, from http://lastingercenter.com/portfolio/algebra-nation-2/

Leite, W. L., Cetin-Berber, D. D., Huggins-Manley, A. C., Collier, Z. K., & Beal, C. R. (2019). The relationship between Algebra Nation usage and high-stakes test performance for struggling students. *Journal of Computer Assisted Learning*, 35(5), 569–581.

Leite, W. L., Jing, Z., Kuang, H., Kim, D., & Huggins-Manley, A. C. (2021). Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–19.

Leite, W.L., Roy, S., Chakraborty, N, Michailidis, G, Huggins-Manley, A. C., D'Mello, S. K., Faradonbeh, M. K. S., Jensen, E., Kuang, H., and Jing, Z. (2022). A novel video recommendation system for algebra, An effectiveness evaluation study. Proceeding of the Learning Analytics and Knowledge (LAK22) Conference.

Lowes, S., Lin, P., & Kinghorn, B. (2015). Exploring the link between online behaviours and course performance in asynchronous online high school courses. *Journal of Learning Analytics*, 2(2), 169–194.

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2021). *Confronting structural inequities in AI for education*. https://arxiv.org/abs/2105.08847

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York, NY: American Council on Education and Macmillan.

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., … Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, 19(2–3), 121–140.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Niaki, S. A., George, C. P., Michailidis, G., & Beal, C. R. (2019). Investigating the Usage patterns of algebra nation tutoring platform. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19* Tempe, AZ, 481–490.

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 1–18.

Poehner, M. E. (2011). Dynamic assessment: Fairness through the prism of mediation. *Assessment in Education, Principles, Policy and Practice*, 18(2), 99–112.

Prieto, L. P., Dlab, M. H., Gutiérrez, I., Abdulwahed, M., & Balid, W. (2011). Orchestrating technology enhanced learning: A literature review and a conceptual framework. *International Journal of Technology Enhanced Learning*, 3(6), 583–598.

Pullin, D., & Haertel, E. (2008). Assessment through the lens of "opportunity to learn." In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp.17–41). Cambridge: Cambridge University Press. Learning in Doing: Social, Cognitive and Computational Perspectives.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.

United States Office of Educational Technology (2017). Reimagining the role of technology in education: 2017 National Education Technology Plan update (*Unnumbered Report*). Washington, DC: Author. Retrieved from https://tech.ed.gov/files/2017/01/NETP17.pdf.

Weller, M. (2007). *Virtual learning environments: Using, choosing and developing your VLE*. New York, NY: Routledge.

Xue K, Huggins-Manley A Corinne, and Leite W. (2021). Semisupervised Learning Method to Adjust Biased Item Difficulty Estimates Caused by Nonignorable Missingness in a Virtual Learning Environment. *Educational and Psychological Measurement*, 001316442110204

Zwick, R. (2019). Fairness in measurement and selection: Statistical, philosophical, and public perspectives. *Educational Measurement: Issues and Practice*, 38(4), 34–41.