Principles for Assessing Adaptive Online Courses

Weiyu Chen
Zoomi Inc.
weiyu.chen@zoomiinc.com

Carlee Joe-Wong
Carnegie Mellon University
carlee.joe-wong@west.cmu.edu

Christopher G. Brinton
Zoomi Inc.
chris.brinton@zoomiinc.com

Liang Zheng
Princeton University
liangz@princeton.edu

Da Cao
Zoomi Inc.
da.cao@zoomiinc.com

ABSTRACT

Adaptive online courses are designed to automatically customize material for different users, typically based on data captured during the course. Assessing the quality of these adaptive courses, however, can be difficult. Traditional assessment methods for (machine) learning algorithms, such as comparison against a ground truth, are often unavailable due to education's unique goal of affecting both internal user knowledge, which cannot be directly measured, as well as external, measurable performance. Traditional metrics for education like quiz scores, on the other hand, do not necessarily capture the adaptive course's ability to present the right material to different users. In this work, we present a mathematical framework for developing scalable, efficiently computable metrics for these courses that can be used by instructors to gauge the efficacy of the adaptation and their course content. Our metric framework takes as input a set of quantities describing user activities in the course, and balances definitions of user consistency and overall efficacy as inferred by the quantity distributions. We support the metric definitions by comparing the results of a comprehensive statistical analysis with a sample metric evaluation on a dataset of roughly 5,000 users from an online chess platform. In doing so, we find that our metrics yield important insights about the course that are embedded in the larger statistical analysis, as well as additional insights into student drop-off rates.

1. INTRODUCTION

Online learning has become a popular way for universities, corporations, and other institutions to offer full classes and certification programs at scale to students outside the traditional campus setting. Yet students in these courses, particularly in those with open enrollment such as Massive Open Online Courses (MOOCs), often exhibit a wide range of backgrounds, degrees of preparedness, and goals. For example, while some may wish to indulge a personal interest, others may wish to refresh their memory of the course material in preparation for a job [11].

Adaptive online courses automatically individualize the content presented to users, and thus hold promise of accommodating student heterogeneity at scale [4]. These course delivery systems may leverage a wide array of measurements to personalize material, such as user performance on assessments and user behavior exhibited while interacting with content and in discussion forums [4]. Both of these forms of data – behavioral and performance – have been shown to be predictive of learning outcomes [2, 6], indicating that they contain information about whether a user's goals have been met. Fully analyzing the different types of user behavior and performance in a course, however, may prove to be overwhelming to an instructor, and may require significant knowledge of statistics in order to properly interpret the analysis.

Thus, it is useful to develop summary metrics that break down insights from user data into a few easily understandable statistics, particularly for large-scale online courses. Such metrics may also allow direct comparisons of the effectiveness of different courses, or of different units within a course. In this work, we propose a mathematical framework and guidelines for such metrics, and demonstrate particular versions of them on a MOOC dataset.

1.1 Research Challenges and Metric Requirements

Education influences both (i) externally observable activity during a course (*e.g.,* performance on quizzes) and (ii) internal user states during and after a course (*e.g.,* knowledge transfer from the course to the workplace) [12]. Any metrics for online (or offline) course efficacy should account for changes in both, but internal changes cannot be observed directly and are often approximated by responses to quiz questions, which are themselves external. For this reason, it is nearly impossible to define a single "ground truth" measure of course quality through conventional learning measurements [24]. Online courses can compensate for this difficulty by collecting many different types of user data, including both user performance as well as user behavioral measurements, which can give a rich picture of how users benefit from the content. At the same time, integrating insights from heterogeneous sources of learning data is itself a challenging task [2].

Adaptive online courses add a further challenge beyond heterogeneous data: unlike non-adaptive courses, they are designed to offer users a consistent experience. A course evaluation criterion must then account for not only overall course *efficacy*, but also its *consistency* across users: such consistency encapsulates how well the adaptation can account for different users and helps to ensure robustness to new, possibly different users joining the system [9]. We therefore identify the following three research challenges:

C1. Incorporating heterogeneous user data: There are at least three types of user data: (i) *behavioral*, *e.g.,* clickstream measurements on course content, (ii) *performance*, both within and external to the course, and (iii) *navigation*, measuring how closely users follow their adaptation path. A metric should be able to combine all or only a subset of this data, and/or other sources, depending on what data is available.

Each of these three measurement types can provide different insights into course efficacy. For instance, some users may obtain high quiz performance while spending a minimal amount of time engaging with the content. This would indicate "success" if a user simply wished to master the course material, but "failure" if he/she also wanted to be intellectually challenged [4]. The navigation data could shed light on this distinction: those who deviate from the recommended path are probably searching for additional material, while those following it are satisfied with the content provided [2]. By combining different types of user measurements, a metric can account for the fact that a low score in one type may not necessarily indicate an ineffective course.

C2. Balancing user consistency with efficacy: Both adaptive and non-adaptive online courses can be evaluated with the user measurements. In either case, high performance may indicate that the course was effective. However, the multiple paths through the material in the case of an adaptive course should also ensure a consistent user experience [4]; high-performing users do not necessarily indicate that the adaptation mechanism succeeded. A metric must thus incorporate consistency as well as an efficacy score.

C3. Online computations: Users generally take weeks or months to complete an online course, which can result in long evaluation cycles if the metric value can only be computed once the course has ended (*e.g.,* with A/B testing or surveys). A metric that can be computed efficiently and regularly updated as users progress through the course is desirable. This online capability would allow instructors to receive feedback as the course progresses, giving them a better opportunity to address weaknesses revealed before the course completes.

1.2 Our Contributions

In this work, we formulate a mathematical framework for metrics that address challenges C1-C3. Our framework takes as input a set of user characteristics derived from observed data of an online course, and we quantify several example characteristics (*e.g.,* path deviation, engagement). To demonstrate our solution, we leverage data from a course that we hosted for Velocity Chess, a popular online chess competition platform that teaches users techniques for playing the game. With this dataset, we compare a comprehensive statistical analysis of the course data with an instance of our metric, and show that the metric provides insights that are difficult to glean from the analysis alone.

More specifically, our work answers the following questions:

(i) How to define metrics that addresses the three challenges? We begin in the next section by presenting our metric framework. To address C2, it includes statistical factors for (i) the consistency of learning characteristics over different users and course units, and (ii) the overall efficacy of the course as indicated by the actual characteristic values. In doing so, to address C1, we account for the fact that different quantities may have different relationships with efficacy; for example, while efficacy is generally linear in quiz performance, *i.e.,* higher performance is a positive indicator, the relationship with time spent is concave, *i.e.,* excessively high time spent indicates confusion. Our metric parameters can also be flexibly chosen to consider different subsets of the quantities, and to induce different priorities on consistency and efficacy. Finally, given the fine-granular timescale at which certain types of learning data are captured, the metric can be computed at any point in the course, addressing C3.

(ii) How to quantify characteristics to be assessed by the metrics? After presenting the metric framework, we derive formulas for several learning quantities that characterize user actions associated with efficacy in a course. We consider three categories of quantities in particular: *behavioral* (*e.g.,* engagement and time spent on content), *performance* (*e.g.,* quiz scores and knowledge transfer), and *navigation* (*e.g.,* deviation from recommendations). While the exact formulas we present for these quantities are specific to the data capture formats of our system, they are readily extensible to other collection mechanisms and content formats too. In performing a statistical analysis of our dataset in terms of these quantities, we observe that (i) while behavior and performance tend to increase throughout the course, they exhibit high variance in different units, and (ii) little correlation exists between most quantities. (i) and (ii) indicate potential room for improvement in terms of efficacy and consistency, respectively.

(iii) How do the insights of the metrics compare to those revealed through full statistical analyses? We then evaluate an instance of our metrics on this dataset, and compare the findings to those of the more comprehensive statistical analysis. Our metric shows that (i) 50% of the users attain less than 16% of the maximum observed metric value, and (ii) a considerable number of users are highly engaged in the course, but performance tends to be low. Both insights are consistent with the findings from the statistical analysis. Additionally, we find that the metric output contains more insight into learner attrition rates than do other course quantities. Overall, we find that our metric can successfully quantify course consistency and effectiveness, giving instructors straightforward statistics that allow them to improve future versions of the course.

We finally review related work on metrics for online courses and recommendation platforms more generally, and then discuss implications and extensions of the work before concluding the paper. In particular, though our metric is designed for adaptive online courses, it is applicable to any personalized recommender system in which multiple signals can give insight into efficacy.

2. OUR COURSE METRIC FRAMEWORK

In this section, we present our metric framework for evaluating adaptive online courses. We first formalize the general architecture of adaptive courses and then specify the combination of consistency and efficacy mathematically.

2.1 Course Architecture and Metric Input

We assume that any adaptive course is organized into a set of units $\mathscr{U}$, with $u \in \mathscr{U}$ denoting a particular unit u. Within each u there can be one or more content files that a user is expected to study, *e.g.,* videos or PDF documents. At the end of u, there may be an assessment quiz consisting of a series of questions. We assume that the course captures user behavior while interacting with the content in u as well as user performance on the corresponding quiz.

Generally speaking, the adaptation logic of the course will recommend for each user a sequence of units $U_r = (u_r(1),...,u_r(t_r))$ to visit, with $u_r(i) \in \mathscr{U}$ denoting the one recommended at time i. This may be different than the actual chronology $U_a = (u_a(1),...,u_a(t_a))$ of the units that the user chooses to visit. The determination of $u_r(i)$ may in general be based on analysis of the user's actions in $u_a(1),...,u_a(i-1)$, including but not limited to their behaviors from interacting with the content, their performance on the quizzes, and potentially sources of data external to the course that are available to the system. Note that certain units may appear multiple times in U_r or U_a, as users may or may be recommended to repeat/revisit one or more units.

2.1.1 Quantities Q

Our metric takes as input a set of characteristics regarding users in the course to be jointly assessed, which we refer to as the set of quantities Q. Each quantity $q \in Q$ can belong to one of at least three categories: behavioral, performance, or navigation, with the latter involving differences between U_a and U_r. The instructor can choose (i) which characteristics are to be used as quantities in Q, and (ii) whether each q is for a particular unit u or across all units in the course. For instance, Q could be just time spent T_u in a single unit u, or $Q = \{T_1,T_2,...,g_1,g_2,...\}$ could be the time spent T_u and assessment grades g_u over all units u in the course.

In this way, the quantities are representative of the (heterogeneous) user feedback to be analyzed by the metric. We discuss the definition of particular quantities for our dataset and data capture system in the next section.

2.2 Distribution-based Metric Framework

The metric framework must use the quantities to determine course consistency and efficacy.

2.2.1 Quantifying Consistency

We incorporate a measure of consistency through the distribution of the quantities Q over users. We construct this distribution over a discretized set of possible quantity combinations, *i.e.,* all feasible combinations of quantities that users could exhibit.

Formally, let $\mathscr{X}$ denote the support of the distribution, *i.e.,* the set of feasible outcomes (note that our empirical samples may cover only a subset of the theoretically feasible outcomes). Further, let $x = (x_1,x_2,...,x_{|Q|}) \in \mathscr{X}$ be a particular point in the support, with x_q being the value of quantity q at this point. The empirical cumulative distribution function (CDF) $F_Q(x)$ over the set of quantities Q is then obtained as $F_Q(x) = \frac{1}{|\mathscr{X}|}\sum_{y\in\mathscr{X}} \mathbb{1}\{y_q \leq x_q \;\; \forall q\}$ along with the associated probability distribution function $f_Q(x)$. Here, $\mathbb{1}$ is the indicator function, and since $f_Q(x)$ is defined over a finite support we have $\sum_{x\in\mathscr{X}} f_Q(x) = 1$.

We wish for the consistency measure to be maximized when the distribution $f_Q(x)$ is concentrated at a single point. To this end, we define the consistency measure

$$M_Q^c(\mathscr{X}) = \sum_{x\in\mathscr{X}} h\left(f_Q(x)\right)$$

where h is a differentiable, strictly convex function on $[0,1]$ with $h(0) = 0$ (no density at x should map to no change in the measure). Strict convexity of h ensures that as density is distributed across more points, the consistency $M_Q^c(\mathscr{X})$ will decrease, a property that we prove formally in our online technical report (see **Proposition 1**) [7]. We could set $h(x) = x^2$, for example.

2.2.2 Combining Efficacy and Consistency

The consistency measure $M_Q^c(\mathscr{X})$ does not carry any information about efficacy: it can be maximized if users concentrate at any point $x \in \mathscr{X}$, regardless of how effective the course is for users at that point. Our metric framework must also incorporate the actual quantity values x_q. To do this, we modify $M_Q^c(\mathscr{X})$ by scaling the $h(f_Q(x))$ by a function of the observed x_q:

$$M_Q^s(\mathscr{X}) = \sum_{x\in\mathscr{X}}\sum_{q\in Q} z_q(x_q)h\left(f_Q(x)\right) \tag{1}$$

We suppose that $z_q(x_q) \geq 0$ for each $x \in \mathscr{X}$. Different choices of the function z_q can then put greater or lesser emphasis on consistency over quantity monotonicity.

Choosing z_q**.** For a given distribution $f_Q(x)$, M_Q^s is monotonically increasing in $z_q(x_q)$ for each x_q. While different values of x for a given individual user would change the estimated distribution f_Q, we suppose that there are sufficiently many users that these changes are small and do not affect M_Q^s's overall monotonicity. The function z_q must therefore be chosen separately for each quantity q to map more effective x_q to a higher $z(x_q)$.

For quantities that are monotonically related to course effectiveness, *e.g.,* quiz performance, we can take $z_q(x) = x$. Most of the quantities q we consider in this work fall into this category, but two of them do not. The first is time spent: a course is ineffective for users who spend an excessively short or long amount of time on it [1,2]. The second is deviation from the adaptive course's recommended path: some deviation from the recommended path can be helpful, particularly to review additional content, but an excessive amount indicates the adaptation is not meeting users' needs. Thus, if q represents either of these quantities, we should take z_q to be a function that initially increases with x_q and then decreases, *e.g.,* a gamma function.

The z_q must also have a component to adjust how much we wish to emphasize consistency compared to monotonicity. For instance, if we define $z_q(x_q) = (1+x_q)^\alpha$ for the parameter $\alpha \in [0,\infty)$, then at $\alpha = 0$ we would only consider consistency ($z_q = 1$). As $\alpha \to \infty$, the z_q term in M_Q^s would dominate the $h(f_Q)$ term, and a larger concentration of users at a more effective point $x \in \mathscr{X}$ would result in a larger marginal increase in M_Q^s, when compared with the increase at a smaller value of α. Thus, for larger values of α, the metric would attain a greater value if a few users have a very positive experience, compared to if all users have a consistent, moderately positive experience. We formally quantify this insight in our online technical report (see **Proposition 2**) [7] by considering, for each value of α, the set of quantity values x for which a consistent experience concentrated at x yields a higher metric value than an inconsistent, uniform distribution of user characteristics over the entire set of feasible quantity values $\mathscr{X}$.

3. DERIVING QUANTITIES FROM DATA

In this section, we derive several specific quantities from learning data that can form the set Q in our metric framework. We do so based on data formats from our course delivery system, considering the case of an adaptive online course we hosted for Velocity Chess, an open chess competition website. We will categorize user activities into three main quantity types: navigation, behavioral, and performance.

While formulating the quantities, we also perform a comprehen-

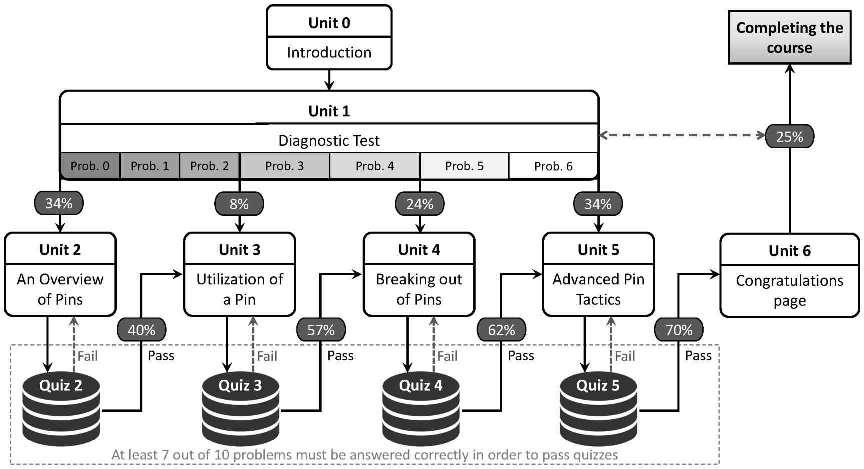


Figure 1: The course consists of seven units: a welcome unit (unit 0), the diagnostic test (unit 1), four core units (units 2-5), and a completion page (unit 6). The adaptation logic is also indicated in the diagram. The percentages indicate the fraction of times that recommendation was made, *e.g.,* in unit 3, a user will answer the quiz and be recommended to advance to unit 4 57% of the time (as opposed to failing the quiz or dropping off before finishing the quiz).

sive statistical analysis of the dataset. In doing so, we make three main findings: (i) many users deviate significantly from their recommended paths, (ii) there is high variability in user behavior and performance, and (iii) user activity and performance tend to increase later in the course. In the next section, we will see that our metric framework also reveals these insights.

Statistical tests. In certain cases, we will run statistical tests to compare distributions of quantities so as to derive qualitative insights into the course efficacy. For these, we will report the p-value (p) and the corresponding test – Wilcoxon Rank Sum (WRS), F-test of Variance, or Pearson correlation [21] – in the description.

3.1 Course Structure and Data Capture

The course we analyze teaches users the Pins strategy for playing chess, from beginner to advanced levels, individualizing the material based on the user's inferred level. It was open to all site users starting in December 2015; we consider the data collected over the one-year time period from December 2015 to 2016, comprising 4,877 enrolled users.

The course architecture and adaptation logic are defined in Figure 1. The content is divided into six units $u = 0, ..., 6$. The core material of the course is contained in Units 2-5, which are of increasing difficulty. Each of these "core units" is comprised of a series of slides and ends with a quiz; after completing the quiz, the course's adaptation logic directs users to a new unit based on their quiz performance. For instance, an average performer may be recommended to proceed to the next unit, but a user who failed the quiz may be asked to repeat that unit. Unit 1 is a diagnostic test that all users take, based on the results of which the adaptation will recommend a core unit to start at.

Clickstream event capture. Each slide in the course is either video-based or text-based. For video slides, the user has a scroll bar to navigate the video, and all `playback` events are captured by the system; these consist of `pause`, `play`, `scrub` (either forward or backward), and `replay` (*i.e.,* starting the slide over), together with the position of the video at which the event occurs. For text slides, there is a single playback event when the user accesses it. In both cases, a slide `change` event is generated when the user moves to a new slide. Slide IDs and UNIX timestamps of all events are also recorded; the IDs include both the previous (immediately before event) and next (immediately after) slides, which differ for `change` events.

The system also records user navigation events independent of particular units: unit `enter` and `exit`, course `login` and `logout`, and application `foreground (fgnd)` and `background (bgnd)`, *i.e.,* when the application is the current active tab on the user's computer. Using these events, we are able to infer a user's navigation between units and their behavior within units. For their quiz performance, we use the `response` events that the system collects after a user answers a question, indicating whether the answer was correct or not.

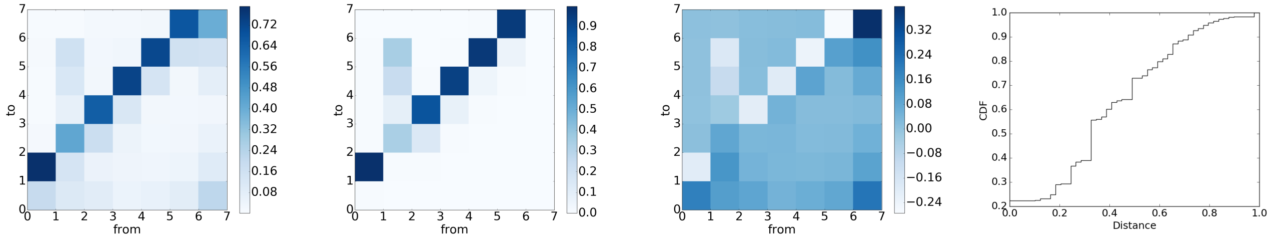
3.2 Quantifying User Navigation

We first investigate user progression through the course units, and use that to define a navigation quantity. Recall that while the system itself generates an adaptation path U_r for each user, the user's chosen path U_a may deviate from the system recommendation. We count a unit as "visited" in U_a if the user spent at least 5 seconds on the material in the unit; time spent on the unit's material is itself a quantity defined in later sections.

Unit-to-unit transitions. 2,186 out of 4,877 users entered the diagnostic test (unit 1) from the introduction (unit 0). For subsequent units, the percentages in Figure 1 summarize the users' recommended paths U_r:

Skill branching: Of the 1,310 users who completed the diagnostic test, the majority (68%) were placed either at the most beginner or the most advanced level. This heterogeneity is common in MOOCs.

Repeating vs. advancing: When placed in core unit u, the fraction recommended to advance to $u+1$ as the next step increased in u (40% to 70%). As users get further through the course, they are more motivated to finish (25% of those who accessed the diagnostic test ended up finishing). Interestingly, very few users are



(a) User navigation.

(b) System recommendation.

(c) Difference (a)−(b).

(d) Distance CDF.

Figure 2: Comparison between (a) user navigation and (b) recommended navigation between units. A point (j,i) in the diagram is the fraction of times unit j was selected while starting on unit i. (c) gives the difference in fractions, illustrating a strong deviation between actual and recommended transitions between units. This is supported by (d) the empirical CDF of the Levenshtein distance d between actual and recommended sequences.

recommended to repeat the core units (less than 3.7% in each case). The remaining users dropped out; we will investigate drop-off further in the next section.

Figures 2a-c show the discrepancies in unit-to-unit transitions between user behavior U_a (a) and system recommendations U_r (b), with the difference between the fractions plotted in (c). In the core units, the vast majority of recommendations are to advance from u to $u+1$, as discussed above. Users' actual paths, on the other hand, are more diverse: there are visibly more repetitions than the system recommends, and also occasional skips back to prior units. Thus, many users likely feel the need for more course content review than is recommended.

Path deviation quantity. We quantify navigation as users' deviation from their recommended paths through the course. To do this, recall the notation $U_r = (u_r(1),...,u_r(t_r))$ and $U_a = (u_a(1),...,u_a(t_a))$. For this course, we always have $t_a \geq t_r$ because navigation can only add steps to the recommended path; users cannot skip units unless recommended. From this, we define the path deviation quantity

$$d = \frac{1}{|U_a|} v(U_a, U_r)$$

where $v(\cdot)$ is the Levenshtein (edit) distance between the two sequences [26]. We choose Levenshtein rather than other distance metrics, *e.g.,* longest common subsequence, because it allows for insertion, deletion, and substitution operations in between strings. In our application, insertion captures users adding additional revising units into U_a from U_r, and substitution captures them choosing to visit different units than those recommended. Division by $|U_a|$ ensures that $d \in [0,1)$.

Figure 2d gives the cumulative distribution function (CDF) of the quantity d over users in the dataset.[1] The mean deviation is 0.36, which can be interpreted as user paths being 36% different from the recommendations on average. On the one extreme, about 22% of users follow the recommendations exactly (*i.e.,* $d = 0$), while on the other hand, 25% of users deviate by 56% or more.

3.3 Quantifying User Behavior

We derive three quantities of user behavior within units: time spent, completion rate, and engagement.

3.3.1 Defining Behavioral Quantities

Let $E = (e_1, ..., e_n)$ be the sequence of n clickstream events generated by a user in the course. For each event e_k, let $s(e_k)$ denote its next slide ID, *i.e.,* the ID immediately after. We write $s \in S_u$ to denote that slide s appears in unit u.

Time spent. Let $t(e_k)$ be the timestamp of event e_k. The time registered for the interval between e_k and e_{k+1} is:

$$T_k = \begin{cases} \min\left(t(e_{k+1}) - t(e_k),\ \tau\right), & \text{if } e_k \neq \texttt{bgnd} \\ 0, & \text{otherwise} \end{cases}$$

In other words, we do not consider time intervals for which the app is in the background, and set the parameter $\tau = 600$ seconds to upper bound the time between actions, capping excessively long intervals when the user likely walked away. From these intervals, the time spent on slide s, T_s, and the time in unit u, T_u, are

$$T_s = \sum_{k:\, s(e_k)=s} T_k, \qquad T_u = \sum_{s \in S_u} T_s,$$

since $s(e_k) = s$ implies that T_k is time spent on s.

Completion rate. Completion of slide s is a binary measure, defined as $R_s = 1$ if $T_s \geq \varepsilon$ and 0 otherwise. We set $\varepsilon = 5$ sec so that if the user spent at least 5 seconds on s it is considered completed. From this, the completion rate of unit u is defined as

$$R_u = \frac{1}{|S_u|} \sum_{s \in S_u} R_s,$$

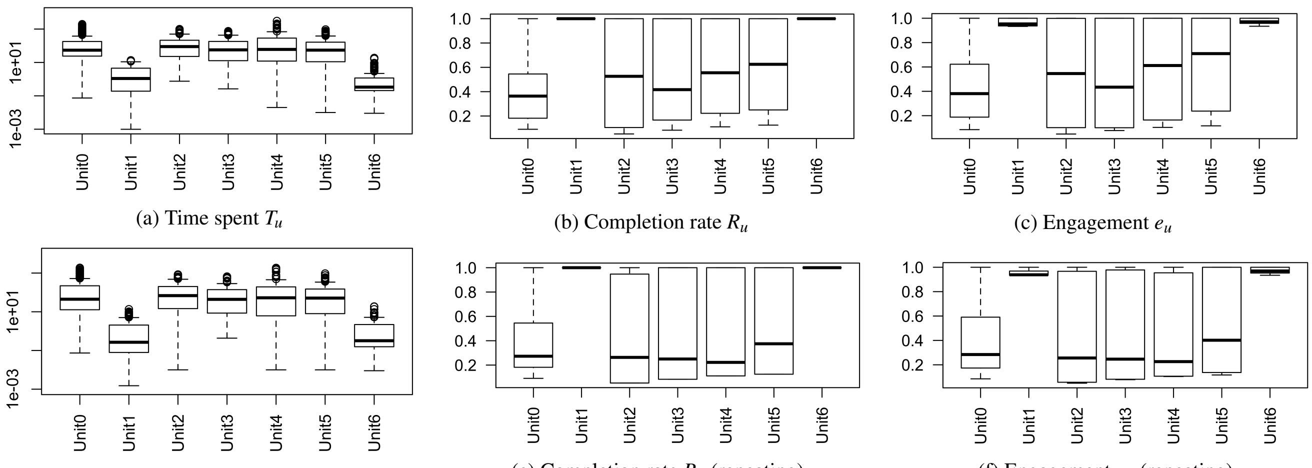
where $|S_u|$ is the number of slides in u. Note that R_u is between 0 (no slides completed) and 1 (all completed).

Engagement. Let $\bar{T}_s$ be the "expected" time spent on slide s. Following the method proposed in [6], we calculate the engagement of a user on unit u as

$$e_u = \min\left(\gamma \times R_u \times \prod_{s \in S_u} \left(\frac{1 + T_s/\bar{T}_s}{2}\right)^{\alpha},\ 1\right).$$

Here, $\alpha \geq 0$ models the diminishing marginal return on time spent, *i.e.,* more time spent on the same slide counts incrementally less towards engagement. The division by 2 makes the computation relative to a user who spends the expected $T_s = \bar{T}_s$ on each slide. $\gamma \in (0,1]$ is a constant that controls the overall spread of the distribution; a user who registers the expected time spent and 100%

[1] In this plot, we only consider users with $|U_a| > 2$, *i.e.,* those who proceeded past the diagnostic test.



(a) Time spent T_u

(b) Completion rate R_u

(c) Engagement e_u

(d) Time spent T_u (repeating)

(e) Completion rate R_u (repeating)

(f) Engagement e_u (repeating)

Figure 3: Distributions of time spent, completion rate, and engagement across units in our dataset. Each quantity is considered both (i) for all user visits to a unit in a-c and (ii) for all visits past the first one (*i.e.,* repeating) in d-f. The core units 2-5 each exhibit significant variation in user behavior.

completion on each slide will have $e_u = \gamma$. By default, we set $\gamma = 1$, $\alpha = 0.1$, and $\bar{T}_s = 60$ sec.[2]

All three behavioral quantities – time spent T_u, completion rate R_u, and engagement e_u – have been defined here on a per-unit basis. We also consider them at a course level to get a complete picture of overall behavior. For these details, see online technical report [7].

3.3.2 *Behavioral Analysis*

Figure 3 gives boxplots of the three behavioral quantities in our dataset, across units. For each quantity, we show behavior over all user visits to units, as well as repeating visits only.

We first observe that *behavior in the core units exhibits high variation in each of the quantities.* The interquartile ranges (IQR) of R_u and e_u are between 0.75 and 0.90, out of a maximum range of 1.0. The ratio of the IQR to the median – a non-parametric coefficient of variation [21] – is larger than 1.2 in each case, up to 4.6 for time spent in unit 4. The IQRs for time spent are up to 275 sec.

Also, *user activity tends to increase in later core units* (WRS $p \leq 0.033$). While time spent (T_u) is reasonably consistent in units 2 to 5 – with medians around 60 sec – completion rate (R_u) and engagement (e_u) both increase considerably from units 3 to 5. In particular, the median R_u rises from 0.42 to 0.63 and the median e_u increases from 0.43 to 0.71. The WRS p-values associated with these changes are significant ($p \leq 0.033$) in each case. Combined with the consistent values of T_u, this implies that users are distributing their time more evenly across slides in later units. This is somewhat surprising because the later material is more challenging, so we would expect certain slides to require more time.

For repetitions, the median t_u drops by < 25 seconds, while R_u and e_u drop more substantially, from 0.17 to 0.39 depending on the unit. The small drops in time spent indicate that users spend a significant amount of time repeating. Coupled with large declines in completion rate, this implies that overall, users are focusing on a more specific set of slides while repeating. Large variations in behavior, however, remain: the third quartiles of R_u and e_u barely move at all.

3.4 Quantifying User Performance

We derive two quantities for user performance: quiz performance and earned virtual currency (called vChips).

3.4.1 *Defining Performance Quantities*

Quiz performance. Let $\mathcal{N}_u = \{n_1, n_2, \ldots\}$ denote the set of questions in the question bank for unit u. Upon a user's lth visit to the quiz for u, they will be given a random subset $\mathcal{N}_u^l \subset \mathcal{N}_u$ of these questions to answer. The number of points earned on the lth visit to u is calculated as $p_u^l = \sum_q p_q^l$, where $p_q^l = 1$ if the user answered question q correctly on the lth attempt, and 0 otherwise. The total points earned on u is then $p_u = \sum_l p_u^l$, and the total points earned in the course is $p_c = \sum_u p_u$. From this, the user's quiz grade on u, g_u, and grade in the course, g_c, are

$$g_u = p_u / N_u, \qquad g_c = p_c / N_c,$$

where $N_u = \sum_l |\mathcal{N}_u^l|$ is the total number of questions answered by user in unit u, and $N_c = \sum_u N_u$ is the total number given to the user in the course. In this way, g_u and g_c are between 0 (no points received) and 1 (all questions answered correctly). Note that, due to question randomization and course adaptivity, $\mathcal{N}_u^l$, N_u, and N_c will vary for each user.

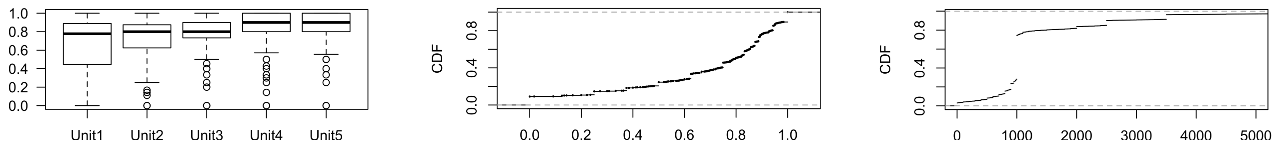
vChips. Velocity Chess awards users vChips[3] – a form of virtual currency – based on their activity and performance on the site. The vChips can be obtained by winning chess games, winning prizes in tournaments, finishing daily challenges, and correctly solving chess puzzles. They can thus measure players' chess skill in practice.

3.4.2 *Performance Analysis*

Figure 4 gives the distributions of the performance quantities g_u, g_c, and vChips. Boxplots of g_u are shown in (a) for each unit that has a quiz, while CDFs of g_c and vChips are given in (b) and (c).

[2] 1 minute is the approximate median of time spent on each slide in the dataset.

[3] https://www.velocitychess.com/faq



(a) Unit quiz performance g_u

(b) Course quiz performance g_c

(c) vChips

Figure 4: Distributions of quiz grades across units, quiz grades across the course, and vChips for users in our dataset. Quiz performance improves in later units, exhibiting significant variation throughout, though less-so than the behaviors in Figure 3. The vChips have a high concentration around 1,000 chips.

Just as user activity increased in later units, we find that *user quiz scores increase further into the course* (WRS $p \leq 0.026$). The median grade in (a) rises monotonically from 0.78 in unit 1 to 0.9 in unit 5. Despite the increase in difficulty, the users reaching later units are likely more knowledgeable and can thus perform better.

We also find that *users' performance is less variable than their behavior* (F-test $p \leq 5.19 \times 10^{-3}$ with the exception of T_u): the IQRs for unit grades g_u range from 0.20 to 0.44, with corresponding IQR-to-median ratios between 0.22 and 0.57. These ratios are smaller than those observed in Figure 3. The vChips have even less variation: with a median of 1,000 and an IQR of 75 chips, the ratio is only 0.075. The vChips have a heavy tail as well, with the mean being 3,271.

3.5 Quantity Correlations

The above analysis indicates that there is high variability in users' behavior and performance quantities unit-by-unit as well as in their vChips and path deviation quantities over the full course. Taken alone, however, any one of these quantities fails to capture the diversity of users taking open online courses. Since our metric framework in Section 2 seeks to aggregate them into an overall measure of efficacy, we also considered the correlation between the different quantities, both between quantities of the same type (Sec. 3.5.1) and between those of different types (Sec. 3.5.2). Overall, we found that most of the quantities exhibit little correlation, i.e., each provides unique information on the diversity of users taking open online courses [11]. In this section, we will present the most interesting of these findings; for the full set of scatterplots and corresponding statistical analysis, see our technical report [7].

Normalizing behavioral quantities. To perform this correlation analysis, we consider each user's quantity values at the course level. To translate the three per-unit behavioral quantities – time spent T_u, completion rate R_u, and engagement e_u – to per-course, we sum all of these quantities over all units of the course for each user,[4] and then normalize over the number of units visited. For completeness, we also considered the number of units suggested by the adaptation algorithm. Formally, let $U_a' \subseteq U_a$ be the set of unique units visited by a user, and $U_r' \subseteq U_r$ be the set of units recommended. The normalized quantities are defined as

$$x_c^a = \frac{1}{|U_a'|}\sum_u x_u, \qquad x_c^r = \frac{1}{|U_r'|}\sum_u x_u,$$

where x_u denotes the quantity (T_u, R_u, or e_u) for unit u, as defined in Section 3. The normalization for x_c^a ensures that the R_c and e_c quantities still lie in $[0,1]$. x_c^r, on the other hand, will become larger than x_c^a when a user takes the initiative of visiting units that were not recommended, i.e., that they could have skipped.

3.5.1 Correlations within Quantity Types

Figure 5 plots the course-level behavioral quantities against one another, normalizing by actual path (x_c^a). We see immediately in Figure 5(a) that there is not a strong relationship between time spent T_c and completion rate R_c, with a Pearson correlation coefficient $r < 0.4$. Those with completion of 100%, in fact, have the highest variation in time spent, perhaps due to them viewing more slides: users' variation in the time spent on each slide would then accumulate over more slides, leading to higher overall variability.

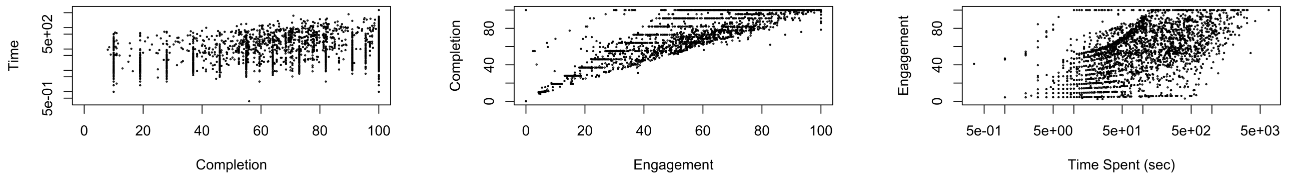
Figure 5(b), on the other hand, shows a strong positive correlation between completion rate and engagement e_c, with $r > 0.95$. This is expected since engagement is defined to be linear in R_u. Specifically, several users have moderate e_c and high R_c: they would have low T_c to pull the engagement level down. Figure 5 shows a positive correlation between e_c and T_c as well, though not as strong, and we can see cases where a low time spent corresponds to a moderate engagement value. Overall, we conclude that *though engagement is a combination of completion rate and time spent, each of the three quantities gives important information on user behavior.*

As for the performance quantities, Figure 6 gives a scatterplot of quiz score g_c against vChips. We see that *vChips and quiz scores are only weakly positively correlated.* The positive association is intuitive, because we would expect those answering the questions correctly to be more skilled in chess and thus to have the potential to win more games. On the other hand, the lack of strength is surprising. There are many uncontrolled factors outside of the course that could affect this, though, such as whether the strategy taught in the course (pins) is useful in a given situation.

3.5.2 Correlations Between Quantity Types

From analysis between quantity types, our key finding is that *the only significant correlation is a positive one between engagement and quiz score*, while the rest of the pairs – distance vs. engagement, vChips vs. time spent, and so on – only have minor associations, if any. This can be seen in Figure 7, which gives scatterplots of selected pairs – vChip and engagement in (a), quiz and engagement in (b), and quiz and distance d in (c) – with behaviors normalized by recommended path (x_c^r). The scatterplot in (b) has a correlation coefficient of $r > 0.75$, meaning that users who complete more slides and/or spend more time on each slide tend to have improved quiz scores. Figure 7(a), on the other hand, shows that *users' vChips are only weakly positively correlated with their behavior*: users with higher engagement do tend to have slightly more vChips, but there are still many instances of low engagement users earning among the most vChips (potentially those with prior knowledge of the pins tactic) and users with high time spent earning the least vChips (potentially those who struggle with the course).

[4] Given the variability between units observed in Section 3, we consider per-unit, per-user quantities in Section 2.



(a) Time spent and completion ($r = 0.395$)

(b) Completion and engagement ($r = 0.959$)

(c) Engagement and time spent ($r = 0.490$)

Figure 5: Scatterplots of the behavioral quantities, normalized by the number of units visited (*i.e.,* x_c^a). The correlation between completion and engagement is strong, but weaker for the other two pairs.

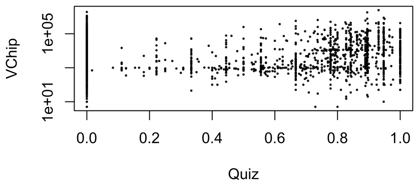


Figure 6: Scatterplot between the performance quantities, vChips and quiz score g_c. There is not a strong correlation between them ($r = 0.138$).

We also found *a weak negative correlation between distance and the behavioral and performance quantities*; the case of distance and quiz score is plotted in Figure 7(c). Users who followed the adaptation algorithm's recommendations, then, have a mild tendency to be more engaged, spend more time, and obtain higher grades than those who deviate from the recommendations. On the other hand, a greater deviation can still lead to lower course activity and grades for some users, and there are different users over the full range of possible completion rates, engagement, and time spent that cover the full range of possible distances. This emphasizes again that the navigation quantity conveys different information than the performance and behavioral quantities.

4. METRIC EVALUATION

The statistical analysis in the previous section revealed that while activity and performance tend to increase further in the course, there is high variability in the quantities overall, and thus room for improvement in consistency and efficacy. In this section, we first perform an evaluation of the course using our proposed metric framework, and show that it also leads to these conclusions. We then consider course drop-off rates, and find that our metric yields better insight into this than do the quantities.

4.1 Course Consistency and Efficacy

Before presenting the results, we first specify particular inputs and parameters of M_Q^s in (1), as well as a sampling procedure to aid in the quantity distribution estimation.

Input quantities Q. The input to M_Q^s is user data on a set of quantities Q. Based on the definitions in the previous section, the full set of quantities Q takes each quantity at the unit-level except distance d and vChips which are only defined over the entire course, *i.e.,* $Q = \{\{e_u, R_u, T_u, g_u \; \forall u\}, d, \text{vChip}\}$. We also consider different subsets of Q in our evaluation, e.g., behavior quantities only.

Functions z_q and h. M_Q^s requires $z_q(x)$ and $h(x)$ for efficacy and consistency. For all metric variations, we take $h(x) = x^2$. We use $z_q(x) = x$ when q is an engagement e_u, completion rate R_u, performance g_u, or vChip quantity, as higher values of these quantities generally indicate a more effective course. We use the gamma distribution $z_q(x) = \frac{1}{\Gamma(k)\theta^k} x^k e^{-\frac{x}{\theta}}$ for the distance d and time spent T_u quantities, reflecting the non-monotonic relationship of these quantities with the course efficacy. We choose θ and k as the squared root of the median value of each quantity, so that g_q attains its maximum value at the median.

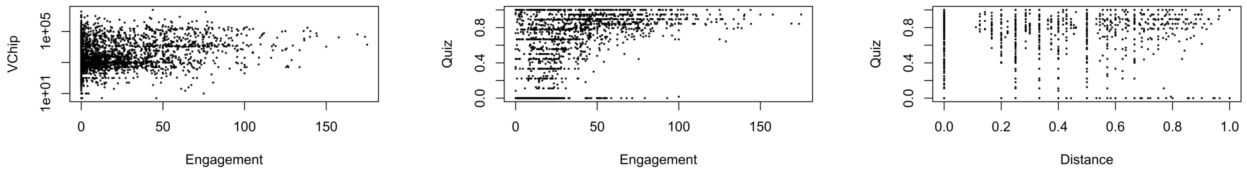
Sampling for $f_Q(x)$. To estimate the distribution $f_Q(x)$ of possible metric values, we first perform random sampling on the realized values of Q to better estimate the properties of the metric output. Similar to bootstrapping [8], for $q \in Q$ we uniformly at random sample non-zero quantity values $x \in x_q$ for each of the users. We take only nonzero values since zero quantity values correspond to inactive users, who may have dropped out of the course or skipped that unit. We take 100 different samples, and combine each with the original dataset to estimate the distribution f_Q and in turn calculate the metric values M_Q^s.

4.1.1 Results and Discussion

Our evaluation results of M_Q^s for the full quantity set as well as subsets are given in Figure 8. Each circle in each distribution plot of Figure 8 represents the metric value from one sample. These plots are the subject of the following discussion.

All quantities. We first consider the metric values over all units and quantities Q. Figure 8a shows the distributions for M_Q^s across samples. We see that many (roughly 50%) of the samplings yield fairly low metric values that are < 1. Considering that roughly 20% of the samples have an output of 6 or higher, meaning that a majority of cases yield less than 17% of the maximum value, this indicates *room for improvement in terms of overall efficacy and consistency*, as we concluded from the statistical analysis. Other samples show clear concentrations around 4 and 6, perhaps due to different quantities concentrating at these values. We also further analyzed the metric in terms of its two constituent pieces – actual quantity values and user consistency – to see whether one had a larger bearing on these low metric values. In doing so, we found that both contribute to low values, confirming room for improvement in both areas; for the corresponding plots, see our online technical report [7].

Behavioral vs. performance quantities. We next compare the metric outputs for the behavior and performance quantities only, in Figures 8b and 8c respectively. Since these quantities reflect different aspects of user activities, we would expect their metric distributions to differ, and we see that this is indeed the case. Also, we

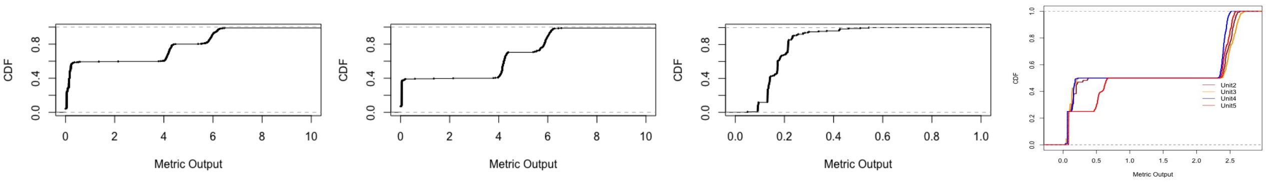


(a) vChip and engagement ($r = 0.129$)

(b) Quiz and engagement ($r = 0.758$)

(c) Distance and quiz ($r = -0.131$)

Figure 7: Scatterplots between selected quantities, with the Pearson correlation coefficient (r) reported for each. Behaviors are normalized by the recommended path (x_c^r). Most of the pairs of quantities exhibit little correlation.



(a) All quantities.

(b) Behavioral quantities.

(c) Performance quantities.

(d) Unit-by-unit quantities.

Figure 8: Distributions of the metric values for 100 different samples. Each circle represents one sample of possible values of Q. (a) is the CDF of M_Q^s considering all quantities. (b) and (c) are the distributions of the metric considering behavioral and performance quantities separately. (d) are distributions of metric values M_Q^s (1) for each unit, in which Q is taken to be each individual quantity. The distributions have a consistent shape for each unit, with over 80% of users experiencing low metric values.

observe that *the metric values are more varied for behavior than they are for performance*, which is consistent with our finding of high variability in behavioral quantities from the statistical analysis. Most users' performance metric values are low, concentrating around 0.2, suggesting poor performance and/or little user consistency. Recalling from Figure 4 that many users performed well on quizzes, this suggests that *these low metric values are likely due to low consistency in scoring*, rather than poor quiz scores. The behavioral metric values, on the other hand, suggest high behavioral quantities and/or high consistency in behavior. The high variability we observed in Figure 3 suggests that *effective behaviors contribute to these higher values*. This conclusion is consistent with the fact that several units show 25% of users obtaining the highest possible engagement and completion rates, whereas time spent is concentrated around its center.

Unit by unit quantities. To analyze differences between units, we also compute the metric over each individual quantity for each core unit. The results are shown in Figure 8d. We see that the distributions are fairly similar for units 2 to 4, exhibiting a fairly wide range of values in each case. As in the distributions over the full course in Figure 8a-c, there is a large concentration of metric values around smaller values, particularly 0. However, the maximum metric values are around 2.5, indicating that some users do have an effective experience in certain units. Indeed, users in unit 5 tend to have the highest values, with roughly 75% of them > 0.5. This is consistent with the conclusion from the statistical analysis that *user activity and performance tend to increase further in the course*.

Overall, these findings indicate that the course is effective at engaging users (Figure 8b), but – at least based on quizzes and vChips – there is room for improvement in teaching them how to play chess (Figure 8c). Given the free and open nature of Velocity Chess's platform, many users likely took the course more out of interest in chess and less out of a desire to memorize chess strategies, which may explain why users' performance is more inconsistent and less indicative of an effective course than their behavior.

4.2 Course Drop-off

We finally validate our metric by comparing it to user drop-off statistics. High drop-off rates are a notorious issue facing open online courses today [3]; we saw in the statistical analysis that our dataset does face this problem particularly in the first three units.

In Table 1, we compare three sets of values across the different units: (i) mean values of behavioral quantities, (ii) metric calculations on the corresponding quantities, and (iii) drop-off percentages, defined as the percentage of users for whom this unit was the furthest visited. Recall that Figure 8d also illustrated the metric values for different units, showing each unit tending to exhibit low values, at least on average.

Overall, we find that *the metrics contain better insight into drop-off than do the behavioral quantities*. Unit 1 experienced a high drop-off while the behavioral quantities e_u and R_u in Units 0 and 1 were fairly high. In particular, on average learners completed almost half of the content in Unit 0 and Unit 1, while almost half of the learners never proceeded past Unit 1. Such drop-off tendencies are difficult to observe from looking at the mean behavioral quantities in Table 1. The metric functions M_Q^s, on the other hand, tell another story; in particular, $M_{R_u}^s$ and $M_{e_u}^s$ in Units 0 and 1 are low when compared to the average values of R_u and e_u. We therefore conclude that when learners are highly likely to drop off, $M_{T_u}^s$ and $M_{R_u}^s$ tend to signal lower quality than do T_u and R_u.

On the other hand, we see that M_Q^s *and the behavioral quantities demonstrate similar trends in the second half of the course, where dropoffs are lower.* Looking at learner behavioral quantities after

	Item	Unit 0	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6
Mean Value	Time spent (T_u)	0.32	0.10	0.05	0.04	0.04	0.08	0.05
	Engagement (e_u)	61.6	46.9	9.82	7.40	8.97	10.44	15.71
	Completion rate (R_u)	41.89	42.16	8.36	6.59	7.63	9.79	14.15
Metric Value	Time spent ($M^s_{T_u}(\mathscr{X})$)	0.04	0.03	0.05	0.05	0.05	0.05	0.03
	Engagement ($M^s_{e_u}(\mathscr{X})$)	21.10	22.31	4.14	4.33	3.86	3.78	5.90
	Completion rate ($M^s_{R_u}(\mathscr{X})$)	4.14	15.12	4.33	4.39	4.07	4.24	5.40
	Drop-off	0.4%	45.0%	14.9%	7.1%	4.0%	5.2%	–

Table 1: Metric comparison with quantities and drop-off rates. The first row of the table entries gives the average learner behavioral quantities, the second row gives the metrics M^s_Q on the corresponding behavioral quantities, and the third gives the drop-off percentages. The low metric values in Units 0 and 1, compared with the corresponding behavioral quantity values, are consistent with these units experiencing high drop-offs.

Unit 2, we observe that T_u, R_u, e_u are generally low, as many learners fail to engage with the course content. The same trend can be observed in the metric values. Interestingly, however, the M^s_Q do tend to increase as the drop-off lessens from Units 3-6, even though our metric was not designed to incorporate this explicitly.

5. RELATED WORK

Learning and content analytics. Recent research in online learning has focused on developing analytics for instructors [2]. Machine learning techniques such as collaborative filtering and probabilistic graphical models have been applied to predict students' abilities to answer questions correctly [17, 23] or their final grades [16, 19]. Other studies have shown that student behaviors display patterns that are significantly associated with learning outcomes [2, 10]. User-content interactions and Social Learning Networks (SLN) have also been used to predict student dropoffs [18, 22], while SPARFA-Trace [13] was developed to track student concept knowledge throughout a course. Few works, however, have studied the efficacy of the course itself, our goal in this work.

Adaptive learning evaluation. Developing course efficacy metrics is particularly important for the growing number of adaptive online courses. For example, MIIC [4] and LS-Plan [14] are all adaptive course delivery platforms that support user- or system-defined individualization across different materials. We can use our metric to improve adaptation algorithms and user experiences. The two most common evaluation mechanisms for adaptive online courses are (i) A/B testing of adaptation versus control group and (ii) user surveys. Although A/B testing [4] allows researchers to test the effect of controlled variations, it is difficult to incorporate additional variables afterwards. Surveys can be used to supplement A/B testing [25], but these rely on user recollections and also cannot be computed at arbitrary points during the course. Our metric framework, in contrast, is easily applicable to different input variables and can be computed at any time during the course.

Online personalization metrics. Substantial amounts of research have been poured into online personalization for applications outside of education, particularly on recommendation systems that predict individual user preferences (see [5] for a survey). Traditionally, these systems have been evaluated with metrics like accuracy and RMSE on a holdout set. Yet these techniques have been criticized as being too distant from the actual user experience [15]. Therefore, newer metrics aim to incorporate factors such as diversity, novelty, and coverage [9,20]. Still, each of these metrics tends to focus on the final results of the prediction without taking into consideration users' prior and subsequent experience with the system. They are also difficult to apply to online courses, which aim to change users' internal knowledge states in ways that are not directly observable.

6. DISCUSSION AND CONCLUSION

We developed a metric framework for adaptive online courses that quantifies both the consistency of users' experiences in the course and the effectiveness of the course across multiple users. To measure effectiveness, we incorporated multiple quantities that describe the full range of user experiences, from their navigation through the adaptive course to their performance on quizzes and external tasks to their interaction with the course material. A statistical analysis of these quantities showed little consistency between different users' experiences and suggested that the course adaptation may not have been effective for many users: many users exhibited poor performance despite spending large amounts of time on the course, and others exhibited high performance but barely engaged with the material. Applying specific instances of our metric to the dataset showed that the metric contained many of the same insights as a statistical analysis, and revealed additional findings consistent with drop-off rates.

A full statistical analysis likely contains more insights than any single metric can provide. Defining a unified metric framework, however, not only allows us to more compactly represent a course's effectiveness, it also allows for direct, quantitative comparisons between different units of a course or even different iterations of a course. This information can then be used by an instructor to improve the material, either in the current or future offerings. While traditional A/B testing requires the instructor to vary one characteristic of the course at a time – which can be inefficient and result in an uneven course experience for different users – our approach enables instructors to estimate the marginal benefits of different interventions, allowing for more rapid and dynamic changes.

Our metric framework is not restricted to adaptive online courses: it can accommodate different quantities that may have distinct relationships to course effectiveness. Indeed, it can even be used for other types of personalized recommendation systems in which multiple quantities can give different insights into the recommendation effectiveness. For instance, users' ratings of a movie on Netflix may contrast with the time spent watching the movie, yielding contradictory information for the recommendation algorithm. Adaptive online courses are, however, perhaps more likely to exhibit such contradictory information than other recommendation settings, and online education presents other unique challenges that require the development of new metrics. The challenges of personalization in different applications motivate the consideration of such metrics more generally.

7. REFERENCES

[1] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton. Studying Learning in the Worldwide Classroom. *Research & Practice in Assessment*, 8, 2013.

[2] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor. Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance. *IEEE Trans. Signal Proc.*, 64(14):3677–3692, 2016.

[3] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong. Learning about social learning in moocs: From statistical analysis to generative model. *IEEE Trans. Learning Technol.*, 7:346–359, 2014.

[4] C. G. Brinton, R. Rill, S. Ha, M. Chiang, R. Smith, and W. Ju. Individualization for Education at Scale: MIIC Design and Preliminary Evaluation. *IEEE Trans. Learning Technol.*, 8(1), 2015.

[5] A. Calero Valdez, M. Ziefle, and K. Verbert. HCI for Recommender Systems: The Past, the Present and the Future. In *Proceedings of the 10th ACM RecSys*, pages 123–126. ACM, 2016.

[6] W. Chen, C. G. Brinton, M. Chiang, and D. Cao. Behavior in Social Learning Networks: Early Detection for Online Short-Courses. In *IEEE INFOCOM*, 2017.

[7] W. Chen, C. Joe-Wong, C. G. Brinton, L. Zheng, and D. Cao. Technical report. https://tinyurl.com/ycm35c3a, 2016.

[8] B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[9] A. Gunawardana and G. Shani. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *Journal of Machine Learning Research*, 10(Dec):2935–2962, 2009.

[10] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video Dropouts and Interaction Peaks in Online Lecture Videos. In *Learning @ Scale*, pages 31–40. ACM, 2014.

[11] R. F. Kizilcec and E. Schneider. Motivation as a Lens to Understand Online Learners: Toward Data-Driven Design with the OLEI Scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):6, 2015.

[12] G. D. Kuh, N. Jankowski, S. O. Ikenberry, and J. Kinzie. Knowing what Students Know and can do: The Current State of Student Learning Outcomes Assessment in US Colleges and Universities. *Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA)*, 2014.

[13] A. S. Lan, C. Studer, and R. G. Baraniuk. Time-varying Learning and Content Analytics via Sparse Factor Analysis. In *SIGKDD*, pages 452–461. ACM, 2014.

[14] C. Limongelli, F. Sciarrone, M. Temperini, and G. Vaste. Adaptive Learning with the LS-Plan System: A Field Evaluation. *IEEE Trans. Learning Technol*, 2(3):203–215, 2009.

[15] S. M. McNee, J. Riedl, and J. A. Konstan. Being Accurate is not Enough: How Accuracy Metrics have Hurt Recommender Systems. In *CHI EA*, pages 1097–1101. ACM, 2006.

[16] Y. Meier, J. Xu, O. Atan, and M. van der Schaar. Predicting Grades. *IEEE Transactions on Signal Processing*, 64(4):959–972, 2016.

[17] Z. A. Pardos and N. T. Heffernan. Using HMMs and Bagged Decision Trees to Leverage Rich Features of User and Skill from an Intelligent Tutoring System Dataset. *JMLR*, 2011.

[18] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue. Modeling and Predicting Learning Behavior in MOOCs. In *ACM WSDM*, pages 93–102, 2016.

[19] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura. Predicting Students' Final Performance from Participation in Online Discussion Forums. *Computers & Education*, 68:458–472, 2013.

[20] A. Said and A. Bellogín. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the 8th ACM RecSys*, pages 129–136. ACM, 2014.

[21] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. crc Press.

[22] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg. Your Click Decides Your Fate. In *ACL EMNLP*, pages 3–14, 2014.

[23] A. Toscher and M. Jahrer. Collaborative Filtering Applied to Educational Data Mining. *KDD Cup*, 2010.

[24] R. M. Wachter. How measurement fails doctors and teachers. The New York Times, 2016.

[25] T.-C. Yang, G.-J. Hwang, and S. J.-H. Yang. Development of an Adaptive Learning System with Multiple Perspectives based on Students' Learning Styles and Cognitive Styles. *Educational Technology & Society*, 16(4):185–200, 2013.

[26] L. Yujian and L. Bo. A Normalized Levenshtein Distance Metric. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007.