Measuring Learning Behaviors Using Data from a Digital Learning Environment

Abstract
This paper describes a partnership-driven approach for developing measures of learning behaviors using event data from a digital learning environment. The approach that we developed and followed involved (1) gathering leaders’, teachers’, and students’ perspectives on learning behaviors; (2) collaboratively analyzing data; (3) using exploratory factor analyses to generate an single score; and (4) conducting explicit model-based tests that assessed the degree to which the single score was correlated with outcomes that were important to all members of the partnership.

Purpose
This paper describes an approach for measuring learning behaviors using data from a digital learning environment. Digital learning environment is a generic phrase that encompasses a variety of technologies, such as games, simulations, intelligent tutoring systems, and learning management systems (LMSs). From a research perspective, an often-cited benefit of digital environments is that actions taken by learners within an environment can be collected and stored in the form of events that capture who did what, when—over time and at-scale (Bienkowski, Means, & Feng, 2013). Granular depictions of learning behaviors collected on large numbers of students offers new opportunities for both researchers and practitioners (Winne, 2017). And while there have been several examples of using event data to quantify students’ learning behaviors and strategies (e.g., Gašević, Jovanović, Pardo, & Dawson, 2017), most of these examples have been developed for research purposes or as proofs of concept (Baker, 2016). In this paper, we describe an approach that went beyond proofs of concept by working directly with teachers and school leaders in the measurement development process.

Theoretical Framework and Methods
Event data generated by digital learning environments introduces different constraints than data typically used for measurement and assessment purposes (Wilson & Scalise, 2016). In typical measurement development, latent, unobservable constructs are clarified over iterative cycles of collecting data (i.e., through a test or questionnaire), analyzing co-variation among data (e.g., factor analyses), and refining potential constructs using available theory. Evidence centered design (ECD; Mislevy, Steinberg, & Almond, 2003) is an example, formal process that adds structure to data-analysis-theory cycles. While approaches like ECD can be applied to event data (e.g., Mislevy, Behrens, DiCerbo, & Levy, 2012), in many situations, researchers and developers must work with data that are already being collected by a digital environment. The challenge, then, is identifying how those data can be used to measure what students know or can do based on pre-set tasks and data tied to those tasks. The data from digital learning environments themselves, therefore, place a considerable constraint on what can be measured.

Given this constraint, what learning behaviors have been identified using event data? Within digital environments and using event data, Beck and Gong (2013) identified an undesirable behavior referred to as wheel spinning, which entails a learner putting forth sustained effort within a learning task without making progress. The flip side of wheel spinning is productive persistence (Kai, Almed, Baker, Heffernan, & Heffernan, 2018) or perseverance (DiCerbo, 2014). These behaviors involve moments where students experience struggle, put forth effort, and make progress using effective strategies (Author, 2016). Other examples include understanding students’ time on task (Kovanović, et al., 2015), conscientiousness (Moore & Shute, 2017), help seeking behaviors (Aleven, Roll, McLaren, & Koedinger, 2016), and efforts to “game the system” (Baker, Corbett, Koedinger, & Wagner, 2004). The benefit of measuring
these behaviors using event data is that they can serve as leading indicators for valued outcomes like passing a course (Author, 2016; 2018). Evidence is beginning to emerge that researchers can effectively operationalize these behaviors and reduce or increase the prevalence of the behavior (Walonoski & Heffernan, 2006) and ultimately improve downstream learning outcomes (Baker, Corbett, Koedinger, Evenson, Roll, Wenger, et al., 2006).

**Collaborative Measurement Development**

In this paper, we position practitioners as the primary consumers to help them in supporting students. As we demonstrate below, the combination of practitioner input and collaborative data analyses, provided the context for grounding choices in specific actions that were operationalized as meaningful behaviors within specific learning tasks.

The partnership was comprised of researchers and teachers and leaders at Charter Management Organization (CMO) operating in the Western United States. At the time of the partnership, which had been in place in various formulations for several years, the CMO operated 10 schools that spanned grades 6-12. One feature of the CMO’s instructional model was a common learning management systems (LMS) that is used by all students to support a large proportion of instructional activity including the completion of playlists that were used in all courses and grades. Playlists are collections of digital resources and assessments; they are widely used across digital learning environments (e.g., Gooru, PowerMyLearning, LRNG, and School of One). Figure 1 presents a visual representation of an individual playlist and the ways in which multiple playlists can be accessed by students within and between courses. The individual playlist on the left-hand side of Figure 1 illustrates how a pre-assessment is the first item and then followed by three resources organized by objectives. Individual resources include a URL to an external website, a link to an external video, and a link to a downloadable PDF file, all of which are for the first objective. As demonstrated on the right-hand side of the image, students can access playlists at various points within and across courses over the school year.

Within the partnering CMO, there is no limit on when a student may start a playlist and no limit on when students can access pre-assessments, summative assessments, or resources within a playlist. Moreover, pre- and summative assessments can be taken multiple times. In order to take a summative assessment, however, students must formally request to do so to a supervising teacher through the LMS. Each time that a student attempts a summative assessment, new items are presented; for each pre-assessment, on the other hand, items remain the same for each attempt. After taking either type of assessment, results are organized by objectives that identify what a student may need to work on and which resources may be most beneficial for a student to later visit.

We analyzed three academic years' worth of data: 2015-16, 2016-17, and 2017-18. Across these three years, we analyzed playlist event data for over 3,000 unique students in grades 6-12 across all core academic courses (e.g., Math, English, Biology). For each academic year, we analyzed approximately 3 million unique events. The overall driving question for the partnership was, *How can event data generated by students attempting to complete playlists be used to measure effective learning behaviors (ELB)?* Put differently, are there ways in which students’ activities can be used to create operationalizations that are meaningful to practitioners and that can help to explain how and why some students are more successful in their efforts to complete playlists? Under this overarching question, we had multiple sub-questions:

- **What are effective learning behaviors from the perspective of teachers, leaders, and students?** To gather input from teachers and leaders, we organized multiple
brainstorming sessions where practitioners identified what they considered to be desirable and undesirable playlist behaviors. To gather input from students, we observed their use of the LMS and interviewed a small sample of students to understanding how they complete playlists.

- **What is the relationship between individual behaviors and students’ performances on proximal learning outcomes?** Teacher, leader, and student perspectives, along with research described previously, all served as the grist for multiple exploratory analyses. A key step in answering this question involved going from over 50 brainstormed behaviors to concrete operationalizations that were positively or negatively related to students’ performances on summative assessments per playlist. Our dominant modeling approach involved multiple cross-classified mixed effects models where an operationalized behavior served as the focal covariate (Raudenbush & Bryk, 2002).

- **Can individual behaviors be turned into a meaningful composite score?** The rationale for developing a single score was to (1) help teachers quickly ascertain the pulse of students’ learning behaviors by offering an initial quantification and (2) to reduce measurement error before it propagates through subsequent modeling activities. Combinations of behaviors were identified playlist-by-playlist across multiple subjects (e.g., English, World History, and Math) and grades (e.g., 6, 9, and 11) by fitting exploratory tetrachoric factor analyses.

- **To what degree is a composite playlist score related to teachers’ coding of playlists?** After developing a single ELB score from brainstormed and operationalized behaviors, we engaged in two analytical approaches explicitly geared toward understanding the construct validity of the combined ELB score. First, we had seven teachers code visual representations of students’ playlist activity; we then correlated teachers’ codes with the combined ELB score.

- **What is the relationship between the ELB score and proximal learning outcomes?** A final validity check involved using the ELB score to assess learning outcomes. To answer these questions, we ran multiple cross-classified mixed effects models.

**Results**

Table 1 provides the operational definitions for ten behaviors that made it from brainstorming activities to exploratory analyses, and during exploratory analyses, demonstrated statistically significant results for each academic year that we had data.

To identify items that work well together, we engaged in multiple exploratory factor analyses. Looking across courses and grades, we examined the possibility of a latent factor that explains variation across multiple behaviors. Using the evaluation metrics of RMSE < .05 and $\chi^2$ model fit of $>.05$, we found that four behaviors consistently performed well together across playlists: Pass PA Before, Before First SA, Assess. Only, and 10 Min. Request (see Table 1). Collectively, we referred to these behaviors as preparing for 1st and subsequent summative assessments, which in the end, became central to the partnership’s understanding of effective learning behaviors. This latent construct, thus, had both statistical support as well as meaningfulness to CMO staff based on the overall process by which the individual measures and ELB score were developed.

After constructing a single ELB score, we subjected the score to two analyses. in an effort to more fully understand the degree to which the ELB score captured that which it was intended to measure. The first test involved seven teachers coding 15 common visual representations of students’ playlist activity. Teachers’ codes represented a proximal measure that was different from our other proximal measure (i.e., summative assessment performance) and was closely
aligned to the focal construct—effective playlist use. Because we had multiple coders, we assessed their interrater reliability by calculating an intraclass correlation (ICC). The specific model that we fit was a two-way, random-effects model that assessed agreement and correlations among raters (Shrout & Fleiss, 1979). Raters had an ICC of .798 (F = 4.95, p < .001). We then calculated an overall correlation between 142 scored playlists not used as part of the inter-rater reliability test (i.e., each one of these playlists was only coded by one rater) and retroactively applied ELB scores to each. We identified an overall correlation of .56 (p < .001), which means that the ELB score was moderately and non-trivially related to teachers’ overall assessment of effective playlist navigation. In total, the results of this first test demonstrates that the ELB score quantifies that which the partnership intended for it to measure to a reasonable degree.

The second set of construct validity tests included putting the ELB score into harm’s way across two model-based tests. The first test assessed the degree to which the ELB score was related to proximal student performance (i.e., average summative assessment performance). Across the two years that we tested this conjecture, a one-unit change in the ELB score was associated with 6.45 and 6.77 positive percentage changes in 2015-16 and 2016-17, respectively, on average summative assessment performances per playlist. These average effects were conditional on different pass rates per playlist, the average number of attempts that students took to pass the summative assessment (i.e., a proxy for assessment difficulty), and the order with which students attempted a playlist per course over the school year.

Significance of the study
The process that the partnership used, over and above the specific findings related to learning outcomes, helps to make clear how event data can be used to identify and understand students’ learning behaviors. The approach we adopted foregrounded construct validity while meeting the needs of teachers and leaders (Penuel, Van Home, Jacobs, & Turner, 2018). In building our evidentiary argument (Mislevy & Haertel, 2006), a high level takeaway from our partnership work is that including practitioners in the process served as an important compliment to the rigorous quantitative analyses carried out in the service of identifying and quantifying effective learning behaviors.

Word Count: 1,994
References
Author (2016)

Author (2018)


Figure 1. Multiple Playlists for One Student Across Three Courses

Note: The student began the school year by accessing a playlist in Course A, which could be the first playlist in the ordered sequence for the course or the last playlist. The next playlist that the hypothetical student accessed was Course B, and the student did not access a playlist in Course C until she accessed two playlists in Courses A and B, respectively.

Table 1. Operational Definitions of Effective Learning Behaviors

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Operationalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before First SA</td>
<td>50% or more of playlist events occurred before first summative assessment attempt</td>
</tr>
<tr>
<td>PA or Resrc. Start</td>
<td>First event on a playlist was a pre-assessment or resource</td>
</tr>
<tr>
<td>Ontime</td>
<td>Earned a passing score on a summative assessment 10 or fewer days after the median student, based on day of the year, passed</td>
</tr>
<tr>
<td>Pass PA Before</td>
<td>Passed a pre-assessment before first summative assessment attempt</td>
</tr>
<tr>
<td>60% Resources</td>
<td>60% or more of playlist events were comprised of resources</td>
</tr>
<tr>
<td>20-Day Gap</td>
<td>Student had a consecutive 20-day gap with no playlist activity; a playlist was deemed to have been started if a student had accessed a pre- or summative assessment or a 3rd resource</td>
</tr>
<tr>
<td>3 Assess. in a Row</td>
<td>Three consecutive events were comprised of three pre-assessments or two pre-assessments and one summative assessment</td>
</tr>
<tr>
<td>Neg. Transition</td>
<td>Student transitioned from a low score on a pre-assessment to a summative assessment as the next event for a given calendar day one or more times</td>
</tr>
<tr>
<td>Assess. Only Day</td>
<td>Student had one or more days of playlist events that were only comprised of pre- or summative assessments</td>
</tr>
<tr>
<td>10 Min. Request</td>
<td>Student requested a new summative assessment 10 or fewer minutes after not passing the same summative assessment</td>
</tr>
</tbody>
</table>