JOURNAL OF LEARNING ANALYTICS

Volume 9(2), 24–37. https://doi.org/10.18608/jla.2022.6751

A Partnership-Based Approach to Operationalizing Learning Behaviours Using Event Data

Andrew E. Krumm[1], Howard T. Everson[2], and Julie Neisler[3]

Abstract

This paper describes a partnership-based approach for analyzing data from a learning management system (LMS) used by students in grades 6–12. The goal of the partnership was to create indicators for the ways in which students navigated digital learning activities, referred to as playlists, that were comprised of resources, pre-assessments, and summative assessments. To develop various indicators, the collaboration gathered school practitioners' perspectives on desirable and undesirable student actions within and across playlists, jointly explored and made sense of LMS data, and examined the relationships between behavioural indicators and outcomes that were important to practitioners. The approach described in this paper is intended to provide an example for future researcher–practitioner collaborations to build upon when seeking to jointly analyze data from digital learning environments. The widespread use of playlists and LMSs in K–12 schools throughout the United States means that the collaborative process described in this paper may have broad applicability to large numbers of digital environments, schools, and collaborations.

Notes for Practice

- Researcher–practitioner collaborations offer an exciting way for schools and technology companies to unlock the potential of data collected and stored within digital learning environments.
- Practitioners and technology developers can play an important role in developing validity evidence for claims and inferences based on behavioural indicators derived from digital learning environment data.
- Complex data sources from digital learning environments can be made useful to practitioners through thoughtfully designed, authentic data analysis and interpretation activities.

Keywords

Researcher–practitioner partnership, learning management systems, learning behaviours and strategies

Submitted: 28/08/19 — **Accepted:** 04/02/22 — **Published:** 25/05/22

Corresponding author [1] Email: aekrumm@umich.edu Address: University of Michigan, Medical School, 221 Victor Vaughan, 1111 E. Catherine St., Ann Arbor, MI 48109-2054, USA. ORCID ID: https://orcid.org/0000-0001-9714-1632

[2] Email: howard.everson@sri.com Address: SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA. ORCID ID: https://orcid.org/0000-0002-8688-1819

[3] Email: jneisler@digitalpromise.org Address: 2955 Campus Dr. Suite 110, San Mateo, CA 94403, USA. ORCID ID: https://orcid.org/0000-0002-6007-9318

1. Introduction

Going from "clicks to constructs" has long been an organizing idea for the field of learning analytics (Knight & Buckingham Shum, 2017). 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 et al., 2012). However, the actual work of translating events — clicks — into meaningful objects of measurement — constructs — is often not straightforward (Winne, 2014). Prior efforts to quantify behaviours like time on task (Kovanović et al., 2015), regularity (Boroujeni et al., 2016), and persistence (DiCerbo, 2014) brought to light multiple complexities that belied the seeming simplicity of these constructs (Gašević et al., 2017). A central concern of both current and past efforts to go from clicks to constructs involves the ways in which digital tasks, which can in principle generate large quantities of data, often produce equivocal data that are open to multiple interpretations as to what generated them (Wilson & Scalise, 2016; Winne, 2017).

In complex organizations — of which schools and universities are prototypical examples — one approach to reduce equivocality in organizational life entails drawing on multiple, diverse sources of expertise using structured sense-making activities, such as cross-functional teams (Page, 2007; Weick, 1995). Equivocality implies that data are possible of multiple meanings (Weick, 1979). Daft and Lengel (1986) argue that when organizational actors face highly equivocal situations, higher-bandwidth forms of interaction like face-to-face communication are needed. In the context of data originating from digital learning environments, higher-bandwidth modes of communication have been shown to benefit data analysis activities, from clarifying the purposes for one or more analyses to aligning purposes with available data (Krumm et al., 2018).

When diverse forms of expertise like different research groups (e.g., Liu et al., 2018), researchers and technology developers (e.g., Coleman, Baker, & Stephenson, 2019), researchers and practitioners (e.g., Zheng et al., 2019), and researchers, practitioners, and technology developers (e.g., Krumm et al., 2021) are brought together, available data can play the role of "boundary object" (Bowker & Star, 1999; Crisan & Munzner, 2019). A central feature of partnerships and projects organized around large, complex datasets is that the data is simultaneously motivating to work with and "unsettled" for both researchers and practitioners (Star, 2010). For researchers, these data can be unsettled because they often do not have requisite contextual knowledge or understandings of specific problems to be solved, despite having analytical skills; for practitioners, these data can be unsettled because they may not have the requisite expertise to model available data, despite having a deep understanding of context and problems to be solved. Strategies for bringing together diverse forms of expertise around large-scale data are beginning to emerge (e.g., Bowers & Krumm, 2021), many of which build on past research describing how diverse teams work with comparatively smaller, less equivocal datasets (e.g., Marsh, 2012).

In this paper, we describe a collaboration between researchers and practitioners in a charter management organization (CMO) operating in the United States. The goal of the collaboration was to understand and measure how students used and navigated *playlists* in a common learning management system (LMS). Students in grades 6–12 across all subject areas completed these playlists as a central learning and assessment activity within the LMS. Playlists, in general, are collections of digital resources and assessments; they are widely used across multiple digital learning environments.

The intended output of exploring how students navigated playlists were formal operationalizations that could be expressed in code, i.e., machine-executable instructions. This "computable knowledge" could then be used by three audiences within the CMO (Friedman & Flynn, 2019, p. 1). The first audience were teachers who wanted to prevent students from falling behind on playlists as well as to prevent students from experiencing unproductive struggle while working on playlists (Beck & Gong, 2013; Beck & Rodrigo, 2014). The second audience were students. Knowledge of how to be a successful student in K–12 schools in the United States is not evenly distributed. Students from dominant cultures have privileged access to what it means to be an "effective" student in systems developed and maintained by the dominant culture (Goldin et al., 2019). Thus, if effective learning behaviours could be operationalized, then all students — at least in principle — could be provided with strategies for navigating an important set of tasks within the CMO. Lastly, the third audience were CMO leaders who were looking for ways of understanding student behaviours that could be monitored to support system-level improvement efforts.

The remainder of this paper is organized as follows. First, we provide an overview of previous partnership-based approaches for working with data and an overview of the LMS used within the CMO. We then describe the methods that the partnership deployed to turn LMS data into computable knowledge for how students navigated playlists. After detailing our multi-step analytical approach, we then describe the results of multiple analyses along with descriptions of the various decisions that the partnership made in jointly analyzing LMS data. Lastly, we describe how collaborations between researchers and practitioners can remove equivocality from digital learning environment data to support teachers, students, and leaders.

2. Partnership-Based Approaches to Working with Data

Research on how partnerships between researchers and practitioners collaboratively make sense of data is beginning to coalesce around two broad purposes: 1) partnerships between researchers and practitioners whose purpose is directed at improving data use within an educational organization, and 2) partnerships who make regular use of data as part of their local improvement projects (Farrell et al., 2018; Piety, 2019). To date, the latter represents the dominant purpose: analyzing data in support of a school's local improvement effort, such as understanding the outcomes of a policy change or assessing the implementation of a new instructional program. For example, Ahn and colleagues (2016), in a partnership with Washington, D.C. Public Schools (DCPS), combined and analyzed data from a digital learning environment and district administrative data system to understand the implementation of that environment and the degree to which students' event data from the environment correlated with learning outcomes. The ways in which Ahn et al. (2016) fed back analyses to DCPS practitioners

ISSN 1929-7750 (online).

to support their decision-making illustrates a different kind of engagement than a more typical "arm's length" evaluation of a district adopting a technology (e.g., Penuel & Gallagher, 2017; Thompson et al., 2017).

While many partnerships use data for improvement purposes, much of this data is newly collected and analyzed by the partnership to address jointly defined purposes. In addition to newly collected data, secondary data from administrative data systems or digital learning environments can be a useful resource within a partnership (Roschelle & Krumm, 2015). However, these data can create unique challenges, such as misalignments between a partnership's goals and the purposes for which the data were originally collected (Figlio et al., 2017). Despite potential challenges, positive cases exist where secondary data have been central to a partnership. The Chicago Consortium on School Research, for example, examined historical data to identify predictors of students likely to drop out of high school (Allensworth & Easton, 2007) and translated those findings into interventions that dramatically decreased dropout rates (Roderick, Kelley-Kemple, Johnson, & Beechum, 2014). The Youth Data Archive (YDA) combined data from multiple youth-serving organizations to inform changes to local policies and programs (McLaughlin & London, 2013; Nelson et al., 2015). Biag (2017), for example, describes a hybrid approach for the ways in which YDA combined historical data with newly collected qualitative data to support family–school partnerships in Redwood City, California.

A central theme across these examples is the role that the partnership played in research and improvement efforts. This theme is also apparent in examples from the field of learning analytics. Ahn and colleagues (2019), for example, outlined their efforts to design data visualizations within the context of a research–practice partnership, whereby their overall approach was based in helping to co-design data visualizations and to support their continued implementation and adaptation over time. Similar to the ways in which Ahn et al. moved beyond simply providing data visualizations to partners, Feng et al. (2016) highlight the importance of researchers jointly defining data analysis goals with practitioners and providing long-term analytical support through collaborative data wrangling and exploration. Clow (2014) outlines a similar, embedded role in his description of the "Data Wranglers" at the Open University.

It is easy to conflate the general, user-centred orientation of the field of learning analytics with the more in-depth collaborations and partnerships described above. Penuel et al. (2015) describe more in-depth collaboration as "joint work," a mode of collaboration where practitioners' "goals for improvement and aspects of the research are defined and evolve through interaction, rather than being planned fully ahead of time or defined by either researchers or practitioners independently of one another" (p. 183). Joint work, therefore, stresses the need for both parties to meaningfully participate in design, development, and implementation activities, i.e., all phases of the work and not just in providing feedback.

The partnership described in this paper used examples from prior data-focused partnerships to structure initial partnership activities around jointly developing research questions as well as collaborative data wrangling and analysis. A key assumption behind forming the collaboration was that joint work could help in reducing the equivocality inherent in the CMO's LMS data. As the collaboration developed over time, joint work included structured activities for setting (and revisiting) goals, defining objects of measurement, rating visualizations of students' event data, and collaborative sensemaking around formative data analyses. In what follows, we describe how these different activities led to quantifications and computable knowledge of students' playlist activity.

3. Overview of Focal Learning Management System

The LMS — around which the partnership organized its collaborative data work — was a key feature of the CMO's overall instructional approach. Each class in grades 6–12 at the CMO provided students with multiple, similarly structured playlists that could be completed at their own pace in addition to their regular in-class activities. While students could work at their own pace, they were expected to complete all playlists for a given course by the end of the school year. Each playlist began with a pre-assessment followed by digital resources organized by objectives, i.e., instructional materials aligned to state standards. Digital resources could include a URL to an external website, a link to an external video, or a link to a downloadable, teacher-created PDF file. Figure 1 presents a visual representation of an individual playlist (left) and the ways in which multiple playlists can be accessed by students within and between courses (right). The individual playlist on the left-hand side of Figure 1 illustrates how a "Pre-Assessment" is the first item followed by three resources organized under "Objective 1." 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. In short, each playlist represents an opportunity to measure what students do as they work to complete summative assessments. Are there students who use a lot of resources? Are there students who use no resources? Do students navigate playlists the same way within and between courses? Across the entire school year, a typical student would complete approximately 30 playlists. Therefore, a student could generate 30 effective playlist navigation measurement opportunities.

ISSN 1929-7750 (online).

Within the CMO, there were few restrictions on when a student could begin working on a playlist; likewise, there were few restrictions as to how students could move within a given playlist. Students, for example, could start a playlist by first accessing the summative assessment to see if they could pass the playlist without using any resources or taking the pre-assessment. To take a summative assessment, however, students needed to formally request to do so from a teacher. Each time that a student attempted a summative assessment, new items were presented; for each pre-assessment, on the other hand, items remained the same for each attempt. Students needed to earn 80% or higher on a summative assessment to complete a playlist; some courses, however, only required that students earn 70%.

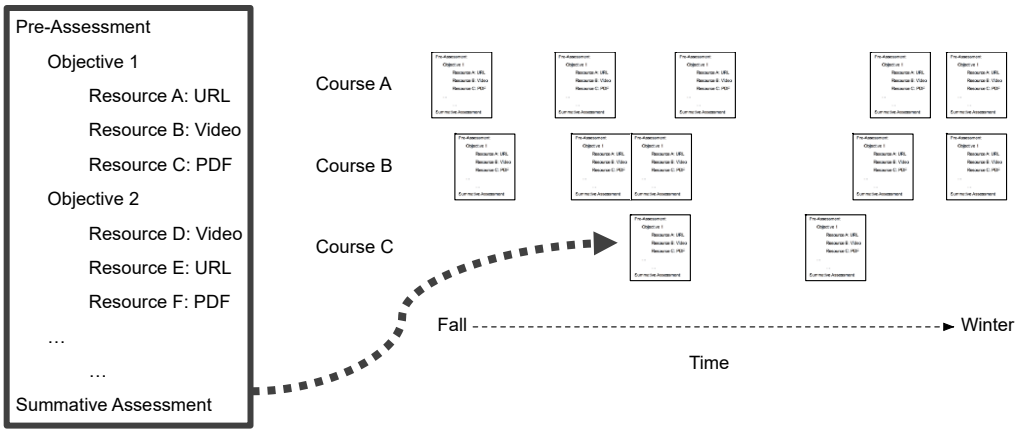

Figure 1. Multiple playlists for one student across three courses.

4. Partnership's Approach to Operationalizing Learning Behaviours

The first two steps in the partnership's approach to quantify learning behaviours within the LMS involved jointly defining the purpose of the partnership and getting to know the available data. Jointly defining the purpose for the partnership entailed co-developing a variety of data explorations and analyses. For the purposes of this paper, we highlight a coherent set of analyses, framed in terms of the following research questions:

1. What are effective learning behaviours from the perspective of teachers and CMO leaders?
2. How much do playlist scores vary across students, courses, and playlists?
3. To what degree do combinations of behaviours relate to teachers' coding of playlists?
4. Do combinations of behaviours strongly and significantly relate to learning outcomes?

Getting to know the data involved exploring data exports from the LMS with CMO leaders. To identify effective learning behaviours, teachers and CMO leaders brainstormed positive and negative learning behaviours. After generating a variety of potential ideas, researchers translated ideas into machine-executable code (RQ 1). Following this step, researchers applied machine-executable definitions to historical datasets that resulted in a score for each playlist. Using these scores, researchers then examined the degree to which scores varied across students, courses, and playlists using random effects models (RQ 2). Following this step, members of the collaboration then correlated scores with two outcomes. The first outcome included ratings that teachers applied to visual representations of students' playlist activity (RQ 3). The second outcome included students' course grades (RQ 4). For RQs 3 and 4, we used LMS data, course grades, and standardized test scores from the 2016–2017 school year, which ended at the start of the partnership's efforts to operationalize playlist learning behaviours.

4.1. Identifying Effective Learning Behaviours [RQ 1]

Beginning in the summer of 2017, four CMO leaders and seven teachers participated in partnership activities. CMO leaders included a curriculum and instruction lead, two research and partnership leads, and one information and technology lead. Teachers included two who were on leave from daily teacher responsibilities to support CMO-level projects and five additional teachers who participated in playlist coding activities described in more detail below. Members of the partnership, in particular the first author and CMO leaders, met on an approximately weekly basis.

Across a series of meetings that took place at the start of the collaboration, teachers and CMO leaders provided their perspectives on effective and ineffective playlist behaviours. Meetings were structured as open forums and were anchored in preliminary analyses of LMS data. During meetings, the first author recorded teachers' and leaders' insights. Working with

teachers and leaders early in the process was intended to create opportunities for practitioner participation, which is central to Penuel et al.'s (2015) definition of joint work.

A key part of the overall operationalization process involved translating individual indicators of a behaviour into a *composite playlist score*. To accomplish this, the collaboration conceived of each brainstormed behaviour as equivalent to a test item that a student could get "correct" or "incorrect" (e.g., Kerr et al., 2016). Using this as our mental model, we created dichotomous indicators for each behaviour and summed each behaviour into an overall score. This approach had several benefits that were weighed against the drawbacks of creating dichotomous indicators. Teachers, for example, wanted to know when to pay attention to a behavioural indicator, and participating teachers' input as well as various exploratory analyses contributed to identifying a cut-off value, or "pay attention" value, for each behaviour. An additional benefit of this approach is that it provided an easy way to combine categorical and continuous behavioural indicators. Moreover, dichotomously scored behaviours were easy to aggregate and disaggregate at the playlist-level, whereby a composite playlist score could direct teachers' attention to a student's overall playlist activity and then broken down into constituent behaviours when providing feedback to a student.

4.2. Variation Across Students, Courses, and Playlists [RQ 2]

An important measurement issue after developing the composite score was understanding the degree to which playlist scores varied across students, courses, and playlists. Understanding this variation was useful for identifying the degree to which students potentially evidence the same behaviours from one playlist to the next, i.e., the degree to which an individual student's composite playlist scores are correlated. To examine the ways in which scores were potentially correlated, we fit two random effects only models using indicators for student, course, and playlist (Raudenbush & Bryk, 2002). We ran this analysis for two populations of students: the first population involved all students in all grades and a second analysis was conducted for students in grades 6–10 who had incoming standardized test scores, which were important to answering RQ 4.

$$Y_{ijk} = \beta_0 + b_{0i} + c_{0j} + d_{0k} + e_{ijk} \quad (1)$$

In Equation 1, Y_{ijk} represents the composite playlist score for student *i* on playlist *j* in course *k*. β_0 is the grand mean with b_{0i}, c_{0j}, and d_{0k} capturing the variance for students, playlists, and courses, respectively. We calculated the intraclass correlation (ICC) for b_{0i}, c_{0j}, and d_{0k} to understand the degree to which composite scores were correlated across student, playlist, and course. The ICC for students, for example, was calculated as follows: $b_{0i} / (b_{0i} + c_{0j} + d_{0k} + e_{ijk})$.

Decomposing variance of composite playlist scores across students, courses, and playlists is similar to a Generalizability study in which researchers are interested in the reliability of an overarching score (Shavelson & Webb, 1991). For RQ 4, it was important to estimate the reliability of an individual student's average composite playlist score within a course. Answering RQ 2, therefore, provided a useful way of understanding the potential measurement error inherent in the score developed as part of RQ 1.

4.3. Relate to Teacher Coding [RQ 3]

After decomposing the variance in the composite playlist score, the collaboration analyzed the degree to which the score correlated with teachers' ratings of students' playlist activity. To collect teachers' ratings, we created a visual representation of students' playlist activity and had teachers rate each image using a web-based survey. We generated 225 images that were then apportioned into seven surveys sent out to teachers to complete on their own time. They were asked to rate on a scale of 1 to 4 the degree to which events in each image were indicative of "effective playlist navigation." In addition to providing a numeric rating, teachers could provide written comments on what they saw in student activity. To complete the coding activity, teachers were given a brief written tutorial on how to complete a rating and how to interpret an image in relation to the events captured by the LMS (see Figure 3). To not overly clutter any visual, especially those with over 100 events, we created a shorthand for different types of LMS events. In Figure 3, a lower-case "t" represented a perfect score on a pre-assessment, a dash "-" signified a resource, a capital "F" represented a failed summative assessment, and a capital "P" signified a passed summative assessment. The day of the school year was placed on the *x*-axis, and the order of events was placed on the *y*-axis with the first event on a given day appearing at the bottom of a column of events and the last event appearing at the top. A reference line "Med. Pass" was provided to show when during the school year the median student earned a passing score on that playlist.

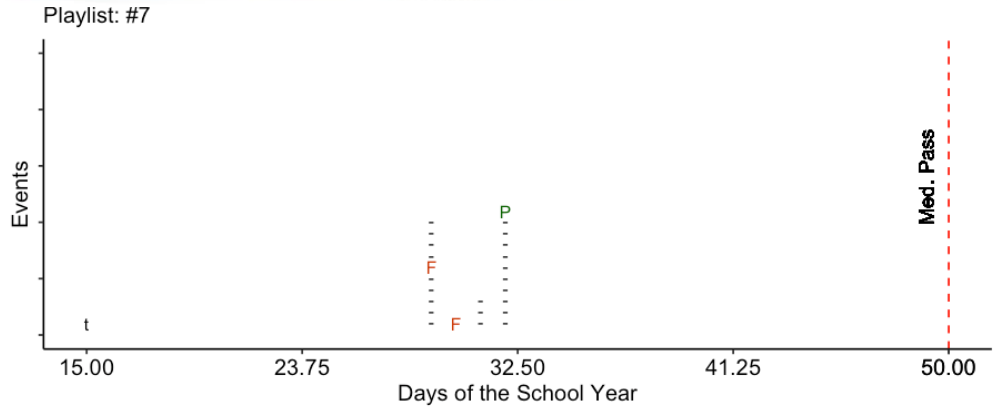

Figure 2. Example events for one playlist. The median student earned a perfect score "t" on a pre-assessment on day 15 of the school year. On day 29, they began working with resources "-", requested but did not pass the summative assessment "F", and ended the day by looking at more resources "-". On day 30, the student requested but did not pass a summative assessment "F", with no other activity on that day; on day 31, the student observed three resources "-"; and on day 32, the student looked at 10 resources "-" before passing a summative assessment attempt "P."

Each teacher coded 50 images; 35 images were unique to the teacher and 15 were shared across teachers to examine interrater reliability. We calculated the intraclass correlation to quantify interrater reliability (Shrout & Fleiss, 1979). Two-way random effects models were used since raters were selected from a broader population of teachers at the CMO and each teacher evaluated the same images. We used both an "average" as well as "single rater" metrics given the newness of the approach and the limited amount of training that teachers received. In calculating ICCs, we addressed "consistency" versus "absolute agreement" to account for systematic error affecting teachers' agreement; it is important to note that "consistency" generally leads to higher ICC estimates (McGraw & Wong, 1996). Each model was fit using the *irr* package (Gamer et al., 2019) in R (R Core Team, 2021). To quantify the relationship between teachers' ratings and composite playlist scores, we used a simple Pearson product-moment correlation coefficient.

4.4. Relationship to Learning Outcomes [RQ 4]

A final question that the partnership sought to answer was the degree to which composite playlist scores that were averaged across playlists per student within a given course correlated with overall course grades. To examine the relationship between playlist scores and course grades, we used a mixed effects model (Raudenbush & Bryk, 2002) and controlled for NWEA Measures of Academic Progress (MAP) test scores from the fall of 2016–2017 (i.e., the beginning of the school year). MAP tests are norm-referenced, computer adaptive assessments used across thousands of school districts throughout the United States. In the CMO, MAP math and reading scores were consistently administered in grades 6–10, and we restricted our analysis to these grades. In a narrow way, the composite playlist score sought to capture non-academic knowledge through the lens of effective playlist navigation skills, i.e., how students engaged in learning, not what they knew. Course grades can be considered a multidimensional assessment of both students' academic knowledge and myriad social-emotional and noncognitive factors (Bowers, 2011). MAP reading (READ) and math (MATH) scores, therefore, served as proxies for students' incoming knowledge.

To answer RQ 4, we created an average composite playlist score (AVGCPS) per student and course. Not all students in grades 6–10 took both READ and MATH, so we ran sensitivity analyses related to AVGCPS where only students who had both were included (reported below) and for READ and MATH alone. Lastly, we included GRADE to adjust for variation across grade levels 6–10. To improve interpretability of the intercept, we group-mean centred AVGCPS, READ, and MATH to have mean 0 and standard deviation 1 within each grade-level. Random effects for student *i* in course *j* accounted for variation in course grades within and between students and courses, respectively. Equation 2 specifies the final model, whereby β_0 is the overall intercept for the model, b_{0i} represents variance across students, c_{0j} represents variance across courses, and e_{ij} is the residual error. To estimate these models, we used the *lme4* package (Bates et al., 2015) in R (R Core Team, 2021).

$$Y_{ij} = \beta_0 + \beta_1(AVGCPS)_{ij} + \beta_2(READ)_i + \beta_3(MATH)_i + \beta_4(GRADE)_{ij} + b_{0i} + c_{0j} + e_{ij} \quad (2)$$

ISSN 1929-7750 (online).

5. Outcomes of Partnership-Based Analyses

5.1. Identifying Effective Learning Behaviours [RQ 1]

Behaviours that were brainstormed by CMO teachers and leaders across multiple meetings ranged from those unique to a given course as well as those potentially generalizable to all courses and grades. Example behaviours included the following:

- Adequately prepare to take summative assessment
- Pass a pre-assessment before taking a summative assessment for the first time
- Avoiding repeated summative assessments
- Maintain a proper balance between resource use and assessments
- Using an appropriate resource following a low pre-assessment score
- Average time between low summative assessment scores
- Start a playlist with a resource or pre-assessment
- Use a high proportion of available resources, if necessary
- Percent of available resources used
- Counts of zero or low percent of points earned on assessments
- Order for when a student took a pre-assessment

Across the above examples, potential behaviours could be either *positive* or *negative* and span units of analysis from individual *key moments* in a playlist, like the events that immediately precede a summative assessment attempt, to combinations of events that encompassed entire playlists. As ideas surfaced, the collaboration developed criteria for selecting final behaviours: Could they be easily operationalized? Were they related to summative assessments scores? Did they remain interpretable after they were operationalized?

Table 1. Operational Definitions of Learning Behaviours

Indicator	Operationalization
1	50% or more of playlist events occurred before first summative assessment attempt
2	First event on a playlist was a pre-assessment or resource
3	Earned a passing score on a summative assessment 10 or fewer days after the median student, based on day of the year, passed
4	Passed a pre-assessment before first summative assessment attempt
5	60% or more of playlist events were comprised of resources
6	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 third resource
7	Three consecutive events were comprised of three pre-assessments or two pre-assessments and one summative assessment
8	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
9	Student had one or more days of playlist events that were only comprised of pre- or summative assessments
10	Student requested a new summative assessment 10 or fewer minutes after not passing the same summative assessment

Through multiple iterations, the partnership decided upon ten behaviours (see Table 1). Indicators 1–5 are all positive, desirable behaviours, and indicators 6–10 are all negative, undesirable behaviours. The logic of using these behaviours to score a playlist in terms of how effective students navigated it is as follows. If a student evidenced a positive behaviour, they would earn 1 point, if they evidenced a negative behaviour, they would lose a point. The highest score a student could earn, therefore, was 5 and the lowest score a student could earn was -5. For example, if a student evidenced indicator 1, 2, and 7 they would earn a score of 1.

5.2. Variation Across Students, Courses, and Playlists [RQ 2]

Table 2 presents the results from the collaboration's effort to understand variation in composite playlist scores across students, courses, and playlists. For Model 1, the average score was 2.13, meaning that students typically evidenced at least two positive behaviours. Importantly, students accounted for 25% of the overall variance in the composite playlist score. This variation

signals that some students had a unique way of navigating playlists, conditional on course and playlist. For example, students who were one standard deviation above average enacted approximately three positive behaviours. For Model 1, the role of course and playlist, combined, accounted for 9% of overall variation, which means that there were some playlists where playlist navigation is correlated (6%) and, to a lesser extent, some courses where playlist navigation is correlated (3%). Most variation (66%), however, was not directly accounted for, which means that despite student-specific behaviors, navigation behaviors varied from playlist to playlist. Variance components were similar between Models 1 and 2. Importantly, the correlation among playlist scores within a given student signalled that the composite score, on average, usefully discriminated one student from another.

Table 2. Variation Across Students, Playlists, and Courses

	Model 1: Grades 6–12			Model 2: Grades 6–10		
Fixed Effects	*B*	SE	*t*-value	*B*	SE	*t*-value
Intercept	2.13	0.06	36.49	2.26	0.06	37.55
Random Effects	Var.	N	% Var.	Var.	N	% Var.
Student	0.80	3,027	25%	0.86	2,298	27%
Playlist	0.19	365	6%	0.18	221	6%
Course	0.09	42	3%	0.05	28	1%
Residual	2.08	144,370	66%	2.09	93,735	66%

Note: Var. = Variance

5.3. Relate to Teacher Coding [RQ 3]

Prior to examining the correlation between the composite playlist score and teachers' individual rating of playlists, we examined the interrater reliability of teachers' coding using an ICC estimate. The ICC for the two-way consistency model for individual raters was .40 (95% CI = .19, .66) and for the mean of raters was .80 (95% CI = .58, .92). One of the seven teachers did not complete their web-based survey, so only six teachers' ratings were included. While rules for interpreting ICCs can vary, values between .40 and .59 can be labelled as "fair," .60–.74 as "good," and .75–1.0 as "excellent" (Cicchetti, 1994). For both ICC estimates, there was a moderate amount of uncertainty as can be seen in the wide confidence intervals (95% CI). The variation in teachers' ratings can also be seen in Figure 3, which shows how many teachers (out of six) gave one of four ratings (*y*-axis) for the 15 common playlists (*x*-axis). Except for playlist 3, every playlist evidenced some variation. Most playlists varied +/- 1 from the modal value, with some, like numbers 4, 10, 13, and 15 varying +/- 2 rating points.

After assessing the agreement among raters, we then calculated an overall correlation between the composite playlist score and teachers' ratings using 209 complete score-rating pairs that were not included in the interrater reliability test. We identified a correlation of .61 ($p < .001$), which means that the composite playlist score was moderately related to teachers' overall assessment of effective playlist navigation. Figure 4 shows the distribution in composite playlists scores (*x*-axis) by teachers' ratings (*y*-axis). Thus, despite the poor to fair agreement among teachers on how to rate effective playlist navigation, there was still a moderate correlation between teachers' ratings and the composite playlist score.

To aid in understanding what teachers attended to as they provided ratings, we examined their written comments. While not every teacher provided comments, available comments signalled that teachers attended to a variety of behaviours when they made their ratings. This finding provides some evidence for the low ICC estimates. Many of the behaviours noted by teachers, however, were captured by indicators contained in the composite playlist score. One teacher, for example, observed the following: "The spread of days here isn't great, but I at least like the usage of resources and pre-assessment prior to [summative] attempts. Interesting the pre-assessment fail right before the [summative] assessment pass — I wonder if the student was just clicking through questions to remind him/herself of content?" These comments address indicators 6, 1, 2, and 8 from Table 1. Another teacher commented that a student "intentionally studied for this [summative assessment]. Followed the process of study, pass [pre-assessment], and take [summative assessment] (and passed!)." Lastly, another teacher observed the following: "Efficient. Used resources. Passed [pre-assessment]. Checked resources after failing [summative assessment] and consequently passed." Across these examples, teachers mentioned gaps in playlist activity (indicator 6), use of resources prior to assessment taking (indicators 2 and 4), success on pre-assessments (indicator 4), and following up a low assessment performance by accessing a resource (indicator 8). While several comments addressed individual behaviours, statements

related to "intentionally" studying, following a "process," and being "efficient" point to some of the challenges of translating a measurement idea into an operational definition. While challenging, the benefit of scoring each playlist with the same ten indicators is that it provides a common, consistent framework, whereby combinations of indicators have the potential for capturing more complex desirable and undesirable behaviours.

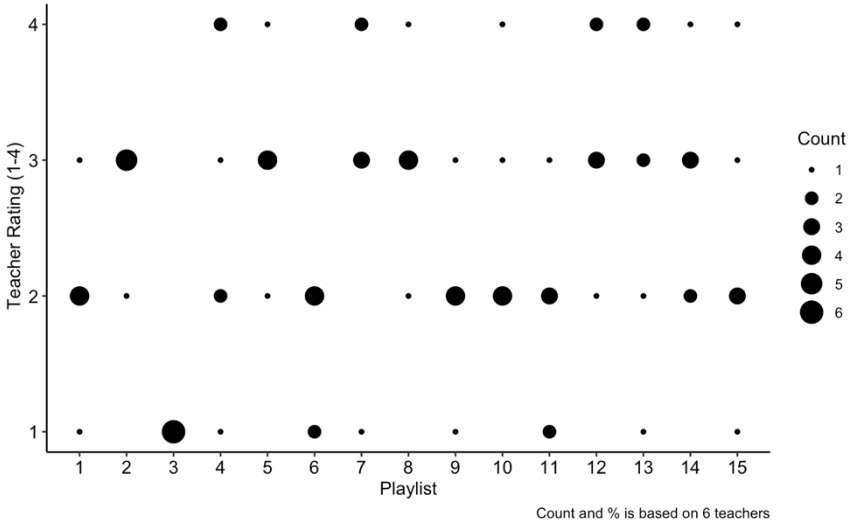

Figure 3. Variation in teacher ratings by playlist.

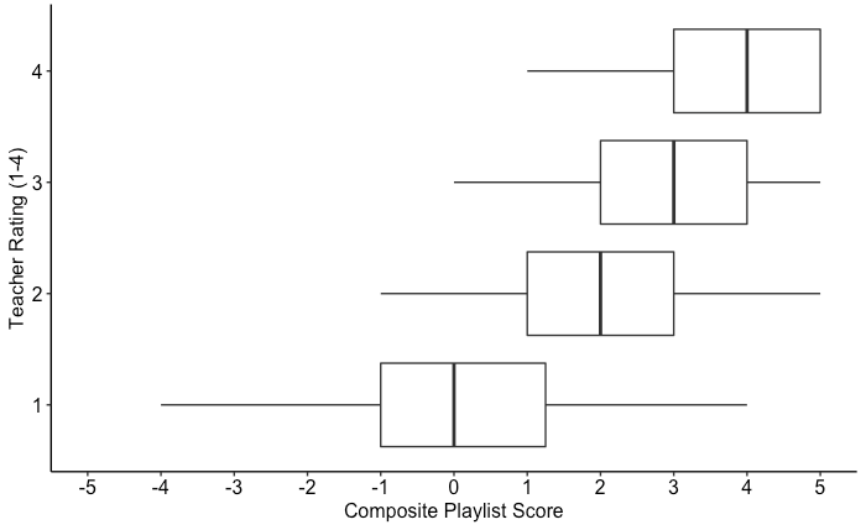

Figure 4. Teacher rating by composite playlist score.

5.4. Relationship to Learning Outcomes [RQ 4]

Table 3 presents our results from the mixed effects models used to test the relationship between students' course grades, operationalized as the percent of points earned, and students' composite playlist score averaged across playlists per course. Model 1, which includes no fixed effects, shows that approximately 76% of the variation in course grades was correlated within students. Model 2 shows the effects on course grades with the addition of multiple control variables: READ, MATH, and GRADE. The inclusion of these variables had a dramatic effect on student and course random effects. READ and MATH helped to reduce student-level variation by 53%. Thus, the correlation among course grades within students remained strong, but MAP scores, on average, helped to explain a large proportion of student-level variation. The control variables across Models 2 and 3 demonstrated strong and significant main effects as well. As can be seen across the levels of GRADE, there were significant, negative differences between some grades and grade 6, the reference category. The main effects for MATH,

ISSN 1929-7750 (online).

in Model 2, were particularly strong: students who started the school year one standard deviation above the mean scored, on average, 4.18 overall course grade percentage points higher. The inclusion of AVGCPS in Model 3 helped to reduce student-level variation while also demonstrating a strong, significant effect: a positive, one standard deviation change was associated with an increase of 1.28 percentage points. Thus, conditional on where students started the year using MAP scores as a proxy for academic knowledge, the ways in which students navigated playlists was significantly related to their final course grades.

Table 3. Relationship to Course Grades

	Model 1			Model 2			Model 3		
Fixed Effects	*B*	SE	*t*-value	*B*	SE	*t*-value	*B*	SE	*t*-value
Intercept	85.33	0.57	150.1	89.39	1.35	66.25	88.75	1.29	68.79
AVGCPS							1.28	0.09	14.28
READ				4.18	0.23	18.13	3.84	0.23	17.05
MATH				2.21	0.23	9.73	1.86	0.22	8.37
GRADE 7				-3.26	1.76	-1.85	-2.73	1.68	-1.62
GRADE 8				-3.90	1.78	-2.19	-3.63	1.70	-2.14
GRADE 9				-4.68	1.82	-2.73	-4.22	1.64	-2.57
GRADE 10				-2.90	1.70	-1.70	-2.33	1.63	-1.43
Random Effects	Var.	N	%	Var.	N	Red.	Var.	N	Red.
Student	80.37	2,293	76%	37.64	1,876	53%	35.28	1,876	56%
Course	6.76	28	6%	5.71	28	16%	5.17	28	24%
Residual	19.20	9,647	18%	17.77	8,273	7%	17.53	8,273	9%

Note: % = % Variance; Red. = % Reduction in variance from Model 1. Grade 6 is reference category for GRADE

6. Discussion

Collaborations between researchers and practitioners offer a promising approach for making sense of event data from digital learning environments. A key starting point for the collaboration described in this paper is that event data generated from a common LMS is capable of multiple meanings. DiCerbo (2017) frames this same point as follows: "It is not clear how to take a log file of activity stream data and identify how the elements should be transformed into observable variables and included in measurement models" (p. 8). Making sense of event data from a common LMS required researchers and practitioners to engage in multiple activities, including defining and revisiting research questions that guided the collaboration's data analysis work. The four jointly developed research questions detailed in this paper, and their accompanying analyses, do not fully address the many analyses that were generated over the life of the partnership. These questions do, however, provide a clear through line and demonstrate the multiple ways in which researchers and practitioners worked together to generate and validate a measure of student behaviours that had local relevance (Gutiérrez & Penuel, 2014).

For RQ 1, teachers and leaders in the CMO brainstormed and helped to operationalize multiple behavioural indicators that could be combined into a composite score. The key challenges faced during this phase of the partnership centred on the translation process of going from an idea to machine-executable instructions (i.e., computable knowledge), which involved developing collaborative approaches and strategies for generating and combining fully operationalized behavioural indicators. For RQ 2, we observed that students engage in similar playlist navigation behaviours across playlists and courses. These analyses demonstrated that when averaged across multiple playlists within a course, the composite playlist score captured a non-trivial amount of signal in an otherwise noisy set of behaviours like navigating digital playlists. For RQ 3, we observed that teachers vary in their appraisal of what constitutes effective paylist navigation. Despite this variation, however, the composite playlist score was still moderately and significantly correlated with teachers' perceptions. Lastly, and for RQ 4, we identified a significant relationship between the ways students navigate playlists and their final course grades, conditional on students' incoming academic knowledge.

Each of these questions provided ways for researchers and practitioners to collaborate around making sense of LMS data. Researchers worked to translate teachers' and leaders' ideas for RQ 1. Likewise, researchers led exploratory analyses and operationalizations that were jointly interpreted by researchers and practitioners for RQ 1. The same researcher-led and jointly interpreted analyses occurred for RQ 2. For RQ 3, multiple teachers participated in a rating activity that provided key validity evidence. Lastly, as with RQs 1 and 2, teachers and leaders helped in making sense of analyses conducted for RQ 4 along with promoting the importance of analyzing the composite score in line with course grades.

7. Conclusion

The collaborative approach described in this paper helped both researchers and practitioners make sense of data from a digital learning environment. Research related to other researcher–practitioner collaborations informed the partnership over time. The overall approach and specific analyses described in this paper are intended to serve as useful examples for other, potential partnerships. The overall approach — comprised of gathering school practitioners' perspectives, jointly exploring data, and jointly examining the relationships between operationalized indicators and outcomes — is intended to provide scaffolding for future partnerships. Moreover, the specific analyses carried out in this collaboration demonstrate that research–practice partnerships can engage in rigorous analytic work that addresses issues of reliability and validity while centring practitioner-driven questions and concerns.

While the collaboration described in this paper provides promising examples for other collaborations to follow, one caution related to our takeaways is that they were generated from a single collaboration (see Krumm et al., 2018 for an effort to synthesize takeaways across multiple partnerships). Another caution related to the collaboration described in this paper is that it did not "close the loop" related to the composite playlist score (Baker & Koedinger, 2018). While leaders and teachers regularly cited the degree to which the collaboration and its resulting analyses supported its improvement efforts, future collaborations should build in strategies to explicitly test the effects of data-informed change ideas. As more and more partnerships develop over time and the scholarship around partnership-driven approaches expands, it is our hope that future collaborations explicitly focus on closing the loop, and that many of the activities described in this paper will become taken for granted ways in which researchers and practitioners jointly work to make sense of event data from digital learning environments.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This project was supported by the partnering Charter Management Organization as a collaboration focused on developing measures of students' learning behaviours to support teachers, leaders, and students.

References

Ahn, H., Beck, A., Rice, J., & Foster, M. (2016). Exploring issues of implementation, equity, and student achievement with educational software in the DC public schools. *AERA Open, 2*(4), 1–10. https://doi.org/10.1177/2332858416667726

Ahn, J., Campos, F., Hays, M., & Digiacomo, D. (2019). Designing in context: Reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics, 6*(2), 70–85. https://doi.org/10.18608/jla.2019.62.5

Allensworth, E. M., & Easton, J. Q. (2007). What matters for staying on track and graduating in Chicago Public Schools. University of Chicago Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/07%20What%20Matters%20Final.pdf

Baker, R. S. & Koedinger, (2018). Towards demonstrating the value of learning analytics for K–12 education. In D. Niemi, R. D. Pea, B. Saxberg, & R. E. Clark (Eds.), *Learning analytics in educatio*n (pp. 49–62). Information Age Publishing.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 431–440). Springer. https://doi.org/10.1007/978-3-642-39112-5_44

Beck, J. E., & Rodrigo, M. M. T. (2014). Understanding wheel-spinning in the context of affective factors. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 162–167). Springer. https://doi.org/10.1007/978-3-319-07221-0_20

Biag, M. (2017). Building a village through data: A research–practice partnership to improve youth outcomes. *School Community Journal*, *27*(1), 9–28.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. U.S. Department of Education.

Boroujeni, M. S., Sharma, K., Kidzinski, L., Lucignano, L., & Dillenbourg, P. (2016). How to quantify student's regularity? In K. Verbert et al. (Eds.), *Proceedings of the 11th European Conference on Technology Enhanced Learning* (EC-TEL 2016), 13–16 September 2016, Lyon, France (pp. 277–291). Lecture Notes in Computer Science, Springer. https://doi.org/10.1007/978-3-319-45153-4_21

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher assigned grades assess in high school. *Educational Research & Evaluation, 17*(3), 141–159. https://doi.org/10.1080/13803611.2011.597112.

Bowers, A. J. & Krumm, A. E. (2021), Supporting the initial work of evidence-based improvement cycles through a data-intensive partnership. *Information and Learning Sciences, 122*(9/10), 629–650. https://doi.org/10.1108/ILS-09-2020-0212

Bowker, G., & Star, S. (1999). Sorting things out: Classification and its consequences. MIT Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–90. https://doi.org/10.1037/1040-3590.6.4.284

Clow, D. (2014). Data wranglers: Human interpreters to help close the feedback loop. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 49–53). ACM Press. https://doi.org/10.1145/2567574.2567603

Crisan, A., & Munzner, T. (2019). Uncovering data landscapes through data reconnaissance and task wrangling. *Proceedings of the 2019 IEEE Conference on Visualization* (VIS 2019) 20–25 October 2019, Vancouver, BC, Canada (pp. 46–50). IEEE Computer Society. https://doi.org/10.1109/VISUAL.2019.8933542

Coleman, C., Baker, R. S., & Stephenson, S. (2019). A better coldstart for early prediction of student at-risk status in new school districts. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (EDM2019), 2–5 July 2019, Montréal, Quebec, Canada (pp. 732–737). International Educational Data Mining Society.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554–571. Retrieved from http://www.jstor.org/stable/2631846

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society*, *17*(1), 17–28. Retrieved from https://www.jstor.org/stable/jeductechsoci.17.1.17

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology, 18*(S1), 7–18.

Farrell, C. C., Davidson, K. L., Repko-Erwin, M. E., Penuel, W. R., Quantz, M., Wong, H., Riedy, R., & Brink, Z. (2018). A descriptive study of the IES researcher–practitioner partnerships in education research program: Final report (Technical Report No. 3). National Center for Research in Policy and Practice.

Feng, M., Krumm, A. E., Bowers, A. J., & Podkul, T. (2016). Elaborating data intensive research methods through researcher–practitioner partnerships. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 540–541). ACM Press. https://doi.org/10.1145/2883851.2883908

Figlio, D., Karbownik, K., & Salvanes, K. (2017). The promise of administrative data in education research. *Education Finance and Policy*, *12*(2), 129–136. https://doi.org/10.1162/EDFP_a_00229

Friedman, C. P., & Flynn, A. J. (2019). Computable knowledge: An imperative for learning health systems. *Learning Health Systems, 3*(4), e10203. https://doi.org/10.1002/lrh2.10203

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. Retrieved from https://CRAN.R-project.org/package=irr

Gašević, D., Jovanović, J., Pardo, A., & Dawson, S. (2017). Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *Journal of Learning Analytics, 4*(2), 113–128. https://doi.org/10.18608/jla.2017.42.10

Goldin, S., O'Neill, M. K., Naik, S. S., & Zaccarelli, F. G. (2019). Supporting student learning practices: Redefining participation and engagement. *The Elementary School Journal, 119*(3), 417–442. https://doi.org/10.1086/701654

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher, 43*(1), 19–23. https://doi.org/10.3102/0013189X13520289.

Kerr, D., Andrews, J. J., & Mislevy, R. J. (2016). The in-task assessment framework for behavioral data. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 472–507). Wiley-Blackwell. https://doi.org/10.1002/9781118956588.ch20

Knight, S., & Buckingham Shum, S. (2017). Theory and learning analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 17–22). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.001

Krumm, A. E., Means, B., & Bienkowski, M. (2018). Learning analytics goes to school: A collaborative approach to improving education. Routledge. https://doi.org/10.4324/9781315650722

Krumm, A. E., Boyce, J., & Everson, H. T. (2021). A collaborative approach to sharing learner event data. *Journal of Learning Analytics*, *8*(2), 73–82. https://doi.org/10.18608/jla.2021.7375

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Does time-on-task matter? Implications for the validity of learning analytics findings. *Journal of Learning Analytics*, *2*(3), 81–110. https://doi.org/10.18608/jla.2015.23.6

Liu, R., Stamper, J. C., & Davenport, J. (2018). A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. *Journal of Learning Analytics*, *5*(1), 41–54. https://doi.org/10.18608/jla.2018.51.4

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*, 1–48. https://doi.org/10.1177/016146811211401106

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30. https://doi.org/10.1037/1082-989X.1.1.30

McLaughlin, M. W., & London, R. A. (2013). *From data to action: A community approach to improving youth outcomes*. Harvard Education Press.

Nelson, I. A., London, R. A., & Stroebel, K. R. (2015). Reinventing the role of the university researcher. *Educational Researcher*, *44*(1), 17–26. https://doi.org/10.3102/0013189X15570387

Page, S. E. (2007). The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press.

Penuel, W. R., Allen, A. R., Coburn, C. E., & Farrell, C. (2015). Conceptualizing research–practice partnerships as joint work at boundaries. *Journal of Education for Students Placed at Risk*, *20*(1–2), 182–197. https://doi.org/10.1080/10824669.2014.988334

Penuel, W. R., & Gallagher, D. (2017). *Creating research–practice partnerships in education*. Harvard Education Press.

Piety, P. J. (2019). Components, infrastructures, and capacity: The question for the impact of actionable data use on P-20 educator practice. *Review of Research in Education, 43*(1), 394–421. https://doi.org/10.3102/0091732x18821116

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models in social and behavioral research: Applications and data analysis methods (2nd ed.). Sage Publications.

Roderick, M., Kelley-Kemple, T., Johnson, D. W., & Beechum, N. O. (2014). *Preventable failure: Improvements in long-term outcomes when high schools focused on the ninth grade year. Research summary*. University of Chicago Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/On-Track%20Validation%20RS.pdf

Roschelle, J., & Krumm, A. E. (2015). Infrastructures for improving learning in information-rich classrooms. In P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrapu, & B. Wasson (Eds.), *Measuring and visualizing learning in the information-rich classroom* (pp. 3–9). Routledge. https://doi.org/10.4324/9781315777979-7

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shrout, P. R., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Star, S. (2010). This is not a boundary object: Reflections of the origin of a concept. *Science, Technology, & Human Values, 35*, 601–617. https://doi.org/10.1177/0162243910377624.

Thompson, K. D., Martinez, M. I., Clinton, C., & Díaz, G. (2017). Considering interest and action: Analyzing types of questions explored by researcher–practitioner partnerships. *Educational Researcher*, *46*(8), 464–473. https://doi.org/10.3102/0013189x17733965

Weick, K. E. (1979). *The social psychology of organizing* (2nd ed.). McGraw-Hill.

Weick, K. E. (1995). *Sensemaking in organizations*. Sage Publications.

Wilson, M., & Scalise, K. (2016). Learning analytics: Negotiating the intersection of measurement technology and information technology. In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An International Compendium of Theory, Research, Practice, and Policy*. Springer. https://doi.org/10.1007/978-3-319-17727-4_44-1

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning, 9*, 229–237. https://doi.org/10.1007/s11409-014-9113-3

Winne, P. H. (2017). Leveraging big data to help each learner and accelerate learning science. *Teachers College Record, 119,* 1–24 https://doi.org/10.1177/016146811711900305

Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, *6*(2), 153–174. https://doi.org/10.18608/jla.2019.62.11