Playing with Data: Investigating a Dashboard to Support Teacher Formative Assessment with Gameplay Data

James Diamond[1], Michelle Cerrone[2], Girlie Delacruz[3]

Roundtable paper prepared for the 2019 Annual Meeting of the American Educational Research Association, Toronto, Canada

Abstract

Teachers must be prepared to use data to monitor student progress and make instructional changes. Teachers are using digital games for formative assessment, but there is little research on how they use the data, or how they can learn to use it to support student learning. We developed a digital dashboard with educative features to scaffold teachers' ability to use gameplay data for formative assessment. We conducted a three-week randomized controlled study with 27 middle school science teachers, comparing teachers who used the educative features to peers without access to those materials. Treatment teachers outperformed comparison teachers on three of six sub-components of a measure of "data literacy for teaching." We discuss the implications for teacher learning and future work.

Keywords

Dashboard; educative curriculum materials; game-based learning; formative assessment; data-driven decision making

Objectives

To meet the needs of all their students, teachers must be prepared to interpret and use data to monitor student progress and make changes to instruction accordingly (Gummer & Mandinach, 2015; Mandinach & Gummer, 2016). That competency goes beyond periodic use of test scores to include using data produced by classroom learning apps and displayed on digital dashboards, which are increasingly part of students' and teachers' daily activities. Digital games are a data source that more teachers are using for formative assessment (Fishman, Riconscente, Snider, Tsai, & Plass, 2014). There is little sustained research on how teachers use gameplay data, however, or how they can learn to use it to support student learning. This study begins to address that gap.

In a three-year project funded by the National Science Foundation, we explored whether a digital teacher-facing dashboard with educative features to support teachers' formative assessment practices would help middle grade science teachers more effectively use data from a video game about argumentation for formative assessment. In a three-week randomized controlled study with 27 middle school science teachers in 22 schools across 12 states (and ~400 students), we compared teachers who used the "educative features" to peers who did not have access to those supplemental materials.

This paper addresses the following research question: Is there promising evidence that teachers who have access to a revised dashboard interface and accompanying educative materials are more proficient in data literacy for teaching than their peers who did not have access to the

[1] Johns Hopkins University School of Education (jdiamo13@jhu.edu)
[2] Center for Children and Technology | Education Development Center (mcerrone@edc.org)
[3] LRNG (girlie@lrng.org)

educative materials? Using a measure of "data literacy for teaching" (Mandinach & Gummer, 2016), our findings indicate that treatment teachers outperformed the comparison group on five of six sub-components of the construct.

Theoretical framework

Game-based learning. Advocates for expanding the role of game-based learning in schools have argued that well-designed video games can enable deep learning by facilitating structured play that is grounded in design principles such as well-ordered problems, situated practice, meaningful feedback, and just-in-time instruction (Gee, 2003; Steinkuehler & Squire, 2014). A body of evidence now generally supports those claims: educational game-based conditions show a moderate advantage over other instructional conditions in science, math, and literacy (National Academies of Sciences, Engineering, and Medicine, 2018).

Video games can function as a form of performance-based assessment when they require players to apply knowledge and skills that they have learned in order to play (Shute, Ke, & Wang, 2017). Performance-based assessment, which entails "the performance of tasks that are valued in their own right" (Linn, Baker, & Dunbar, 1991, p. 15), helps educators to use authentic tasks (i.e., those that emulate situations where a skill might be used outside of the testing scenario) to observe students use the skills they have learned. When integrated with other classroom learning activities, games that have been aligned to learning objectives can be useful tools for teachers to assess and help build student competencies.

Data-driven decision making. Data-driven decision making is a systematic process for collecting and interpreting data to guide decisions about policy and instruction; it is a type of formative assessment when teachers use information systematically to inform their teaching practices (Mandinach, 2012). Mandinach, Gummer, and Muller (2011) and Mandinach (2012) noted that teachers have used data from classroom quizzes and observations for a long time, often informally, to gauge student progress. Most teachers are not trained to use data systematically during their pre-service education, however (Mandinach, et al., 2011), and their access to professional development to build data literacy skills is typically limited (Means, Chen, DeBarger, & Padilla, 2011).

Formative assessment. Formative assessment involves teacher practices for gathering data about student learning and making changes to instruction—it is assessment for learning, rather than assessment of learning (Bennett, 2011). To conduct formative assessments skillfully, teacher must have domain knowledge, pedagogical content knowledge, knowledge about students' previous learning, and assessment literacy (Heritage, 2007). There are comparatively few studies that document what teachers do when they review, interpret, and make decisions about student data (Little, 2012). This is especially true in the case of gameplay data. While Fishman, Riconscente, Snider, Tsai, and Plass (2014) found that teachers do use games for formative assessment, they did not analyze the quality of those practices.

Educative curriculum materials. Curriculum materials can be created to support student learning *and* to improve teachers' content knowledge and pedagogical content knowledge (Ball & Cohen, 1996; Davis & Krajcik, 2005; Krajcik & Delen, 2017). These materials can help teachers improve their instructional practices (Beyer & Davis, 2009; McNeill, 2009). Recommendations from previous research figured heavily in the design of this study's educative dashboard to help teachers' build competencies in data-driven decision making and formative assessment (Davis & Krajcik, 2005; Davis, et al., 2014). While the dashboard is not a curriculum, it includes materials for use with students and to inform teachers' content knowledge and pedagogical content knowledge about argumentation in science.

Methods

Research design. We conducted a clustered randomized impact study, using mixed methods. Participants in the treatment and control groups had access to a video game about argumentation, a data dashboard, a supplemental mini-unit for classroom instruction, and supporting materials to differentiate instruction. Teachers in the treatment group also had access to our intervention, which is an additional "layer" of educative materials that are accessed from the dashboard.

The educative materials were designed to deepen teachers' understanding of gameplay data and build their formative assessment competency in four ways: strengthening their content and pedagogical content knowledge about argumentation; helping them see how the game operationalized four basic argumentation skills; contextualizing the data on the dashboard and connecting the skills as practiced in the game to argumentation in the real world; and helping them differentiate instruction based on student progress. Teachers used a supplemental mini-unit on "Energy and Argumentation" over the course of 15 consecutive days, alternating between game play and lesson days so they could review student game play data and make instructional decisions based on the data. Teachers could adapt the lessons as needed.

Group assignments. We assigned teachers to the treatment or control groups using a block assignment design, establishing equivalency on years of experience and whether the teacher had previously taught argumentation (see Table .5). We assigned teachers on a rolling basis to allow different teacher start dates. As a proxy for random assignment, we assigned the teachers in each block to either the treatment or control group on an alternating basis, based on their date of study enrollment.

Table .5. Teacher demographics

	Control (n=12)	**Treatment** (n=15)
Grade level		
Grade 7	4	7
Grade 8	7	7
Grades 7 & 8	1	1
Experience		
Years Teaching	15.1	12.9
Experience teaching argumentation	8	10

Data sources

Our measures included: (1) a data literacy for teachers assessment; (2) a pre- and post-assessment of student knowledge of argumentation skills (Osborne, Henderson, MacPherson, Szu, & Wild, n.d.); (3) teacher think-aloud sessions; (4) a weekly teacher implementation log; (5) the A-GAMES survey, to gather data on teachers' use of games in the classroom as well as their use of games for formative assessment (Fishman, Plass, & Riconscente, 2013); and (6) an end-of-study teacher survey about teachers' experiences implementing the study materials. We focus on two measures below.

Data Literacy for Teaching. To assess teacher proficiency in data literacy for teaching, we created a mock set of data for teachers to review. Teachers reviewed the data during a two-part timed Web conference interview, which served as the teacher assessment. During part 1, teachers thought aloud as they examined dashboard data. During part 2, a researcher asked the teacher seven prompts. We administered the assessment upon completion of the three-week intervention.

We developed a rubric to score the assessment, drawing from Mandinach and Gummer's (2016) data literacy for teaching framework. Our rubric consists of six data literacy components: (1) articulate inferences and conclusions (AIC); (2) probe for causality (PFC); (3) determine next instructional steps (NIS); (4) understand data in the context of gameplay (UDC); (5) synthesize diverse data (SDD); and (6) assess patterns (AP). We used teachers' responses in part 2 for our analysis because part 1 was open-ended, and part 2 included standardized prompts, yielding a more accurate picture of teacher proficiency through structured questions designed to elicit behaviors of interest.

Scoring. We used an expert-validated rubric to code the interview transcripts by identifying segments of text where teachers exhibited one or more the sub-components. We then scored each segment using the same rubric. We also created a composite score of the sub-components. To ensure reliability, we double-coded and double-scored the coding and scoring of the segments, resolving differences by discussion.

Results

Below we describe the results of the descriptive and inferential analyses of teacher outcomes. Given the small sample size and exploratory nature of this work, the results are meant to test whether there is preliminary evidence of promise that our intervention supports teachers in developing data literacy. Future work will test the intervention on a larger and more representative sample.

Descriptive analysis

Descriptive statistics. We calculated descriptive statistics (Table 1) on all variables that we hypothesized would mediate outcomes related to our intervention, including use of gameplay for formative assessment, general formative assessment practices, and years teaching. We then conducted independent sample t-tests to identify baseline differences between the treatment and control groups across the outcome variables (Table 2).

Results of the t-tests indicate there are statistical differences between treatment and control groups, at or below the .05 alpha level for the total composite score, AIC, NIS, AP, and below the .10 alpha level on UDC and SDD. On average, teachers in the treatment group outperformed teachers in the control group for the total composite score for all but one sub-component (PFC).

Analytical model

We conducted analyses of covariance (ANCOVAs) to test whether there were mean differences between the treatment and control groups using teacher scores on the data literacy components as the dependent variables and a dummy variable indicating condition (control=0; treatment=1) as the fixed factor. To control for teacher-level characteristics, we included these covariates: experience using formative assessment, experience with game-based learning, and total years teaching. We ran seven ANCOVAs—the first looking at the intervention effect on the total composite score, and the rest looking at the intervention effect on the six individual sub-components. This enabled us to determine whether the intervention supported some aspects of data literacy more than others. To adjust for teacher-level characteristics, we included three covariates in our models: years teaching, frequency of using gameplay data for formative assessment, and frequency of using information from formative assessment.

Rather than conducting a fully powered analysis to determine statistical significance, we focused on the magnitude of the effect size to determine whether it was "substantively important," according to What Works Clearinghouse (2010), which would indicate that there was promising evidence that our intervention supported teachers' data literacy, and ultimately their proficiency with formative assessment practices.

Tables 3–9 display results from each ANCOVA. Adjusting for teacher characteristics, results indicate statistically significant differences between the treatment and control group for the total composite score F (1, 21)=8.993, p<.01, ES=.3. We also found differences for three of the six data literacy components:

1. Articulating Inferences and Conclusions [F (1, 21)=8.859, p<.01, ES=0.3];
2. Next Instructional Steps [F (1, 21)=4.729, p<.05, ES=0.2]; and
3. Assessing Patterns [F (1,21)=4.418, p<.05, ES=0.2].

Although the group differences were not statistically significant for Understanding Data in the Context of Gameplay and Synthesizing Diverse Data, the effect sizes suggest small, but substantive differences: F (1, 21)=2.332, p=.142, ES=.1 for UDC and F (1, 21)=2.591, p=.122, ES=0.11 for SDD. We did not find mean differences for Probing for Causality, F (1, 21)=.187, p=.67, ES=.009. Likewise, we did not find any interaction effects.

Discussion

The results of these analyses provide evidence of promise that access to educative materials supports teacher data literacy. Our analyses indicate that, on average, teachers in the treatment group outperformed their peers in the control group on the overall measure of data literacy. We also found mean differences for all but one of the components, and statistically significant differences in three of the components: AIC, NIS, and AP. Additionally, group differences for UDC and SDD approach statistical significance. Additional research is needed to better understand the strength of these relationships and the generalizability of our results.

Significance

This study is an early contribution to a field in need of empirical investigations of how teachers use gameplay data for formative assessment, and how they can be supported in this process. Our findings demonstrate promising evidence that digital dashboards that are designed intentionally with educative features can help classroom teachers learn to use gameplay data more effectively for monitoring student learning and making decisions about instruction. Just as curriculum materials have traditionally been the center of teaching practice and proved effective as tools to support learning new content and instructional practices, well-designed dashboards can be tools to help teachers use data more effectively for formative assessment. This is particularly promising in the area of educational games: well-designed games are good tools for generating performance-based data, and teachers who are prepared to use that data formatively will provide high-quality game-based learning experiences for their students. In future work, we will refine the assessment rubric, investigate how different features/elements had an impact on teacher outcomes, and determine whether these types of dashboards generalize to other games.

Study Limitations

Although this study demonstrates evidence of promise, its generalizability is limited due to a small sample size that is not representative of the larger population of middle school science classrooms. A larger-scale study is needed to establish the role of digital dashboards with educative features on the larger population and to parse out the influence of individual educative features on teachers' data-driven decision making. Similarly, we need to explore if the impact we found can be generalized to other digital games. Importantly, although we engaged in a rigorous process to develop the teacher assessment and establish content validity, we hope to continue to refine the assessment as we learn more about the construct of data-driven decision making.

References

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational researcher, 25* (9), 6–14.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy, and practice, 18*(1), 5–25.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, *86*(1), 79–122.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational researcher*, *34*(3), 3–14.

Davis, E., Palincsar, A. S., Arias, A. M., Bismack, A. S., Marulis, L., & Iwashyna, S. (2014). Designing educative curriculum materials: A theoretically and empirically driven process. *Harvard Educational Review*, *84*(1), 24–52.

Fishman, B., Riconscente, M., Snider, R., Tsai, T., & Plass, J. (2014). Empowering Educators: Supporting Student Progress in the Classroom with Digital Games. Ann Arbor: University of Michigan. Retrieved http://gamesandlearning.umich.edu/a-games/.

Gee, J.P. (2003). *What video games have to teach us about learning and literacy.* New York: Palgrave Macmillan.

Gummer, E., & Mandinach, E. (2015). Building a conceptual framework for data literacy. *Teachers college record*, *117*(4).

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi delta kappan*, *89*(2), 140–145.

Krajcik, J., & Delen, I. (2017) The benefits and limitations of educative curriculum materials, *Journal of science teacher education, 28*(1), 1–10, DOI: 10.1080/1046560X.2017.1279470

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher, 20*(8), 15–21.

Little, J. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American journal of education, 118*(2), 143–166.

Mandinach, E. B., & Gummer, E. S. (2016). Every teacher should succeed with data literacy. *Phi delta kappan*, *97*(8), 43-46.

Mandinach, E. B., & Gummer, E. S. (2016). Data literacy for educators: Making it count in teacher preparation and practice. Teachers College Press.

Mandinach, E. B., Gummer, E. S., & Muller, R. (2011). The complexities of integrating data-driven decision making into professional preparation in schools of education: It's harder than you think. Alexandria, VA/Portland, OR/Washington, DC: CNA Education/Education Northwest/WestEd.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). Teachers' ability to use data to inform instruction: challenges and supports. Washington, DC: Office of Planning, Evaluation and Policy Development, US Department of Education.

Osborne, J., Henderson, J.B., MacPherson, A., Szu, E., & Wild, A. (n.d.). Assessments of Argumentation in Science. Retrieved from http://scientificargumentation.stanford.edu/assessments/

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Switzerland: Springer International Publishing.

Steinkuehler, C., & Squire, K. (2014). Videogames and learning. In R.K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed.) (pp. 377–396). New York: Cambridge University Press.

Table 1. Descriptive Statistics

	Min	Max	Mean	Standard Deviation
Articulate Inferences and Conclusions (AIC)	1	4	2.42	0.76
Probe for Causality (PFC)	2	4	3.12	0.82
Next Instructional Steps (NIS)	2	4	2.69	0.74
Understand Data in the Context of Gameplay (UDC)	1	4	2.27	1.40
Assess Patterns (AP)	2	4	3.15	0.97
Synthesize Diverse Data (SDD)	1	4	2.38	0.98
Total composite score	10	22	16.04	3.40
Years experience	3	27	14.12	7.81
Total time spent on dashboard (in minutes)	44	266	131.19	60.40
Game-based assessment composite	1	4.2	1.96	0.93
Formative assessment component	3	5.0	4.43	0.53

Table 2. Results of t-test and descriptive statistics: treatment vs. control

	Treatment Condition				95% CI for Mean Difference df=24		
	Control (*n* = 11)		Treatment (*n* = 15)				
	M	SD	M	SD	Lower	Upper	t
Total composite score	14.09	2.81	17.47	3.14	-5.84	-0.91	-2.83*
Articulate Inferences and Conclusions	2.09	0.54	2.67	0.82	-1.16	0.01	-2.03**
Probe for Causality	3.18	0.87	3.07	0.80	-0.57	0.80	0.35
Next Instructional Step	2.36	0.67	2.93	0.70	-1.14	0.00	-2.08**
Understand Data in the Context of Gameplay	1.73	1.27	2.67	1.40	-2.04	0.16	-1.76**
Synthesize Diverse Data	2.00	0.78	2.67	1.05	-1.44	0.11	-1.78*
Assess Patterns	2.73	0.91	3.47	0.92	-1.49	0.01	-2.05**
Years experience	15.45	8.20	13.00	7.59	-4.11	9.02	0.78
Total time spent on dashboard (in minutes)	119.00	60.38	140.13	60.89	-70.85	28.58	-0.88
Game-based assessment composite	1.93	0.87	1.99	1.01	-0.84	0.72	-0.16
Formative assessment composite	4.47	0.49	4.40	0.57	-0.36	0.51	0.36

*p<.05, **p<.01, ***p<.001

Table 3. ANCOVA: total composite score by condition and teacher characteristics

	Data Literacy: Total Composite Score			
	Mean	SD		n
Treatment	17.47	3.14		15
Control	14.09	2.81		11
Source	SS	df	MS	F
Treatment	74.85	1	74.85	8.99**
Formative assessment experience	.976	1	.976	0.74
Game-based learning experience	32.76	1	32.76	3.94
Years teaching	0.03	1	.03	0.00
Error	174.78	21	174.78	

R^2 = .40 *p<.05, **p<.01, ***p<.001

Table 4. ANCOVA: Articulating Inferences and Conclusions (AIC) by condition and teacher characteristics

	Data Literacy: Articulating Inferences and Conclusions			
	Mean	SD		n
Treatment	2.67	.816		15
Control	2.09	.539		11
Source	SS	df	MS	F
Treatment	2.98	1	2.98	8.86**
Formative assessment experience	0.22	1	0.22	0.64
Game-based learning experience	4.74	1	4.74	14.10**
Years teaching	1.45	1	1.45	4.32
Error	7.06	21	0.34	

R^2 = .51 *p<.05, **p<.01, ***p<.001

Table 5. ANCOVA: Probing for Causality (PFC) by condition and teacher characteristics

	Data Literacy: Probing for Causality			
	Mean	SD	n	
Treatment	3.07	.799	15	
Control	3.18	.874	11	
Source	SS	df	MS	F
Treatment	0.12	1	0.12	0.19
Formative assessment experience	1.43	1	1.43	2.27
Game-based learning experience	0.07	1	0.07	0.10
Years teaching	0.88	1	0.88	1.39
Error	13.23	21	0.63	

$R^2 = .21$

Table 6. ANCOVA: Next Instructional Steps (NIS) by condition and teacher characteristics

	Data Literacy: Next Instructional Steps			
	Mean	SD	n	
Treatment	2.93	.704	15	
Control	2.36	.674	11	
Source	SS	df	MS	F
Treatment	2.41	1	2.41	4.73*
Formative assessment experience	0.12	1	0.12	0.24
Game-based learning experience	0.58	1	0.58	1.14
Years teaching	0.35	1	0.35	0.68
Error	10.70	21	0.51	

$R^2 = .21$ *p<.05, **p<.01, ***p<.001

Table 7. ANCOVA: Understanding Data in the Context of Gameplay (UDC) by condition and teacher characteristics

	Data Literacy: Understand Data in the Context of Gameplay		
	Mean	SD	n
Treatment	2.67	1.40	15
Control	1.73	1.27	11

Source	SS	df	MS	F
Treatment	3.98	1	3.98	2.33
Formative assessment experience	0.16	1	0.16	3.00
Game-based learning experience	1.86	1	1.86	1.09
Years teaching	3.90	1.	3.90	2.29
Error	35.81	21	1.71	

$R^2 = .27$

Table 8. ANCOVA: Synthesize Diverse Data (SDD) by condition and teacher characteristics

	Data Literacy: Synthesize Diverse Data		
	Mean	SD	n
Treatment	2.67	1.05	15
Control	2.00	0.78	11

Source	SS	df	MS	F
Treatment	2.62	1	2.62	2.58
Formative assessment experience	0.04	1	0.04	1.86
Game-based learning experience	0.03	1	0.03	0.03
Years teaching	0.00	0.00	0.00	0.00
Error	21.27	21	1.01	

$R^2 = .12$

Table 9. ANCOVA: Assess Patterns (AP) by Condition and Teacher Characteristics

	Data Literacy: Assess Patterns		
	Mean	SD	n
Treatment	3.47	0.92	15
Control	2.73	0.91	11

Source	SS	df	MS	F
Treatment	4.42	1	4.42	5.42*
Formative assessment experience	1.12	1	1.1.2	1.38
Game-based learning experience	1.76	1	1.76	2.16
Years teaching	0.90	1	0.90	1.10
Error	17.11	21	0.82	

$R^2 = .27$ *p<.05, **p<.01, ***p<.001