Computers & Education 130 (2019) 1–12

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Digital games as sources for science analogies: Learning about energy through play

Wendy Martin[a,*], Megan Silander[a], Sarah Rutter[b]

[a] Education Development Center, 96 Morton St. 7th Floor, New York, NY, 10014, USA
[b] Mount Sinai School of Medicine, 1 Gustave L. Levy Pl, New York, NY, 10029, USA

ARTICLE INFO

Keywords:
Improving classroom teaching
Pedagogical issues
Teaching/Learning strategies

ABSTRACT

Many research studies have focused on how game designs informed by learning theory can benefit students; fewer explore what needs to happen during instruction for students to translate what they learn in games into disciplinary knowledge. In this study, researchers examined whether instructional techniques that support visual mapping of analogies can help students translate game learning to science learning. The three digital games used in the study were designed to have images and actions that are analogous to photosynthesis, electricity, and heat transfer, topics that all relate to the larger crosscutting concept of energy transfer. We hypothesized that these digital games could serve as compelling, visual sources for analogies, but that teachers would need to explicitly map the relationships between the game visuals and the target science concepts for students to learn those concepts. Our study compared student learning outcomes from an intervention that combined digital gameplay with analogy mapping to outcomes from an intervention that used the same digital games without analogy mapping. We found that students who experienced analogy mapping learned more as measured by assessments of electricity and energy transfer. Findings from this study provide practical recommendations for integrating digital games into classroom instruction, as well as insights into how explicit analogical mapping supports sense-making and measurable progress toward accurate conceptual knowledge of complex scientific phenomena.

1. Introduction

Many researchers have argued that digital games have the potential to help students learn (Gee, 2014; Young et al., 2012; Young & Slota, 2017) and more recent meta-analyses have shown positive effects of digital games in comparison to non-game conditions under certain instructional conditions (Clark, Tanner-Smith, & Killingsworth, 2016; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). However, much of the research on educational games has focused on how aspects of game *design* are associated with learning (Callies, Gravel, Beaudry, & Basque, 2017; Clark et al., 2016; Law & Chen, 2016; Sun, Chen, & Chu, 2018; Young & Slota, 2017) in either lab studies or highly-controlled classroom studies of game interventions. There is less evidence about the specific kinds of instructional scaffolds that can enable teachers in real-world classroom settings to make effective use of digital games (Bell & Gresalfi, 2017a; Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011). Some researchers have found that student learning improves when teachers have debrief discussions that connect gameplay to content covered in the classroom (Rowe, Asbell-Clarke, Bardar, Kasman, & MacEachern, 2014). This study was designed to build knowledge about how to structure those kinds of discussions

[*] Corresponding author. Education Development Center, 96 Morton St. 7th Floor, New York, NY, 10014, USA.
E-mail addresses: wmartin@edc.org (W. Martin), msilander@edc.org (M. Silander), Sarah.Rutter@mssm.edu (S. Rutter).

https://doi.org/10.1016/j.compedu.2018.11.002
Received 2 March 2018; Received in revised form 6 November 2018; Accepted 8 November 2018
Available online 10 November 2018
0360-1315/

so that teachers can help students transfer what they learn in games into an understanding of the concepts covered in classroom instruction. We explored whether students who played digital games and then engaged in discussions that drew analogies between the games and science concepts they were being taught would learn more than students who played the games but did not engage in those discussions. Though this study focused on middle school science, analogy mapping is not specific to any one discipline. Therefore, the findings have broad implications for classroom integration of digital games in any content area.

1.1. *Theoretical framework*

Educational researchers and developers have long seen the promise of digital games for creating educational experiences that inspire students to learn subject matter content with the same enthusiasm that they learn rules and strategies in well-designed commercial digital games (Gee, 2014; Lu, Buday, Thompson, & Baranowski, 2016; Travis, 2017; Young et al., 2012). However, for the skills and strategies gained from games to be transferred to a target disciplinary domain, that knowledge needs to be translated by a skillful interpreter, using scaffolded instruction, (Culp. Martin. Clements, & Presser, 2015; Barzalai & Blau, 2014; Clark et al., 2016; Tamim et al., 2011; Wouters et al., 2013). It is important for teachers to guide students through the process of reflecting on the strategies used in the gameplay and connecting what they learned from the game to real-world concepts and content (Rowe, Bardar, Asbell-Clarke, Shane-Simpson, & Roberts, 2016; Van Der Meij, Leemkuil, & Li, 2013). These “bridging discussions” enable teachers to build on the implicit learning that occurs when players engage with game mechanics and challenges, so that students can gain an explicit understanding of the scientific principles they are intended to embody (Hattie & Yates, 2013; Rowe et al., 2014).

Analogies are a common strategy to facilitate thinking and the development of new conceptual knowledge in science (Braasch & Goldman, 2010; Cooperrider, Gentner, & Goldin-Meadow, 2017), as well as in other domains, such as mathematics (DeWolf, Son, Bassok, & Holyoak, 2017; Richland & Begolli, 2016). Through analogical thinking, students are able to build knowledge about new or less-well-known topics, based on their prior knowledge (Gentner & Smith, 2012; Holyoak & Thagard, 1996). Analogies can be very effective for helping students understand concepts that are hard to comprehend because they enable students to connect a concept with which they are familiar (the analogy *source*) to concepts they are learning in the classroom (the analogy *target*) (Vendetti, Matlen, Richland, & Bunge, 2015). Digital games provide a particularly compelling source analogy because they can be designed to ensure clear systematic correspondences in processes and structures to the target concept (Habgood & Ainsworth, 2011). Having a whole class play a digital game prior to instruction can be an effective means to ensure that students with different background knowledge develop a shared understanding of the analogy source from which to build new knowledge (Reese, Tabachnick, & Kosko, 2015).

Researchers have identified specific ways teachers can integrate analogies into instruction that are particularly effective for learning (Gentner & Maravilla, 2018; Richland & Begolli, 2016). First, the connection between the analogy source and the target should not be limited to surface-level or physical similarities (the *source* and the *target* look the same or share a common feature), but rather they should demonstrate comparable relationships, structures, or processes (certain things relate to each other in the *source* in a way that is similar to how things relate to each other in the *target*). This study explored the use of a pedagogical approach called analogy mapping as a way to transfer the knowledge gained from a game to an understanding of science concepts (Gentner & Maravilla, 2018; Richland & Begolli, 2016; Richland, Zur, & Holyoak, 2007) by identifying underlying relationships and the parallels and discrepancies between contexts (Asmuth & Gentner, 2017; Vendetti et al., 2015). Certain techniques such as visuals, prompts, comparative gestures, and guiding questions can help students explicitly map the relationships between the source and the target. Digital games can be particularly useful in this mapping process because they contain visuals and actions that the players/students are likely to remember and may be excited to discuss with peers and teachers after gameplay (Gee, 2013; Hayes & Gee, 2012). Comparing and contrasting the key objects and relations within one system that is well-known, like a game that has been played repeatedly by a whole class, to those in a second system that is less well known, like a new science concept, can support the kind of conceptual leaps necessary for understanding complex scientific phenomena (Jaeger, Taylor, & Wiley, 2016; Richland & Simms, 2015; Vendetti et al., 2015). In this case, digital games serve as sources for relational analogies that teachers map to the target concepts during instruction. Using analogy mapping, teachers guide students to apply knowledge of the initial familiar source—a digital game—to reason and develop new inferences about a target science concept.

1.2. *Research question*

This study was designed to build knowledge about the instructional supports that are needed to make the most effective use of digital games for science learning. We explored one method of operationalizing the use of analogies in instruction, providing specific instructional supports for the explicit mapping of the visualizations in digital games to similar processes in the target science concepts. Our study focused on the following research question:

- Do students who play digital science games and then engage in discussions that map analogies from the games to the concepts they address demonstrate greater understanding of those concepts than students who play the games but do not engage in those discussions?

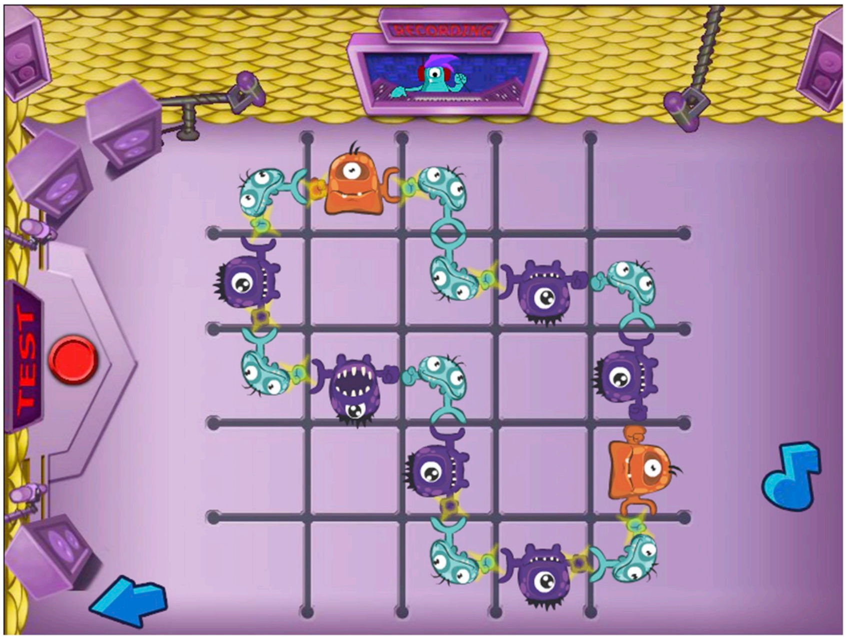

Fig. 1. Image of monsters holding hands in Monster Music.

2. Methods

2.1. Intervention

We conducted a matched comparison group study to answer this question. We created instructional sequences on topics included in the Next Generation Science Standards for middle-school science—photosynthesis, heat transfer, and electricity—that address the crosscutting concept of energy transfer. Each of the sequences included a digital game based on the one of the three topics. (Our organization created these games under a prior grant from the U.S. Department of Education's Institute of Education Sciences.) We selected these topics because they are ones that students find challenging and about which they often hold misconceptions (Andersson, 1990; Chi, 2008; Ozay & Oztas, 2003; Smith, diSessa, & Roschelle, 1994; Squires, Driver, & Rushworth, 1994). Each game was designed to have images and actions that are analogous to the target science concepts. For example, in the electricity game, the goal is to create music CDs to energize the exhausted citizens of a town called Harmonia. To create the CDs, players have to complete an alignment puzzle by turning monster musicians on a grid in various directions to make them hold hands (Fig. 1). Each monster has one open and one closed hand, which represent positive and negative charges. Before they are properly aligned, each monster makes a cacophonous noise. When they all are aligned, or holding hands, the cacophony changes to a flow of music, just as electricity flows from positively to negatively charged atoms through a closed circuit. Players have to do this puzzle multiple times during gameplay to achieve the game goals, so the visualization becomes very familiar to them. In the photosynthesis game, players have to make glucose chains to power a robot through a cave. In the heat transfer game, the player is a zookeeper in outer space who has to alter the temperature of moving globules that are sent to warm an egg enough to hatch it. All of the games and related instructional materials were piloted tested in classrooms and revised based on feedback from teachers and students (Martin, Silander, Culp, Brunner, & Parris, in press).

In this study, we designed treatment and comparison interventions that were very similar, except that the treatment condition used analogy mapping techniques while the comparison condition used an alternate set of instructional techniques (See Table 1 for a comparison of the two sequences.). Both conditions included three instructional sequences in the same order. Students in the comparison condition played the games for the same amount of time as did those in the treatment condition. Keeping the gameplay time the same between the two conditions allowed us to isolate the relationship between the different pedagogical approaches and student learning from the gameplay itself. Exit tickets—brief formative assessments given to students at the end of a class to check for understanding—were administered after both groups played each of the games. Our analysis of student responses on these demonstrated that there was no difference in the level of understanding of the game goals and core game challenge. Each sequence also required that teachers use the same topic-related instructional PowerPoints (photosynthesis, heat transfer, and electricity) and the same topic-related activities. In both the treatment and the comparison conditions, teachers scaffolded learning by having students draw upon existing knowledge from which to build an understanding of the target concept, using small-group and whole-group discussions. However, the conditions differed in the kind of prior knowledge that students were asked to draw upon.

For the treatment group, we incorporated research-based cognitive supports for analogies, based on a schema of cues to focus attention on relational similarity developed by Richland et al. (2007) for teachers' use of analogies. The specific techniques include:

Table 1
Instructional sequences for treatment and comparison GroupsClasses

	Treatment	Comparison
Part 1	30 min of gameplay	30 min of gameplay
Part 2	Activating prior knowledge: Classroom discussion of what students did in the game and the strategies they used to move through the level, using the game debrief PowerPoint	Activating prior knowledge: Prior-knowledge activity where students write down what they already know about the topic, with small-groups discussions to answer a topic-related question from *Uncovering Student Ideas in Science* prompt and share out answers
Part 3	Teacher-led presentation on the topic, using researcher-designed instructional PowerPoint	Teacher-led presentation on the topic using researcher-designed instructional PowerPoint
Part 4	Topic-related hands-on classroom activity	Topic-related hands-on classroom activity
Part 5	Play game again for 15 min	Play game again for 15 min
Part 6	Connecting prior knowledge to target concept: Classroom discussion of how the digital game relates to the topic concepts using the analogy mapping PowerPoint	Connecting prior knowledge to target concept: Small-group discussions about whether the group members think the original answer they gave to the topic-related question in *Uncovering Student Ideas in Science* is correct or should be revised. Groups share their decision with class.

1. Using a familiar source analog to compare to the target analog being taught;
2. Presenting the source analog visually;
3. Keeping the source analog visible to learners during comparison with the target;
4. Using spatial cues to highlight the alignment between corresponding elements of the source and target;
5. Using hand or arm gestures that signal an intended comparison; and
6. Using mental imagery or visualizations.

The treatment condition included instructional time for students to have a discussion about the game after gameplay to help clarify the source of the analogy and ensure that students developed a full and shared understanding of the mechanisms in the game that underpin the analogy. This “debrief PowerPoint” included animations from the game to prompt students and teachers about the important processes embedded in the game. Then, after instruction on the topic, a topic-related activity, and a second round of gameplay, the treatment teachers presented the “analogy mapping PowerPoint,” which showed animations of the games next to images of the target scientific concept, and led a discussion using analogy mapping techniques that allowed the teacher and students to draw analogies between the game and the science concept, using visual supports. The PowerPoint presentation prompted discussions of the analogy by presenting students and teachers with questions about various specific mechanisms of the game and how the game and the target concept are similar (or different), followed by a summary of each similarity or difference between the target and source analog. For example, the PowerPoint presentation for the electricity game includes the following questions along with animations showing this aspect of the game. Students share their ideas, then they go to the slide that has the answer.

All of the text below is shown on a series of slides that *have an animation of the monsters in the grid alongside an image of an electrical circuit.*

> Question slide: What do the monsters need to start making music?
>
> Answer slide: To start making music, all of the monsters need to have their hands connected correctly.
>
> Question slide: How is that related to what you learned about current electricity?
>
> Answer slide: As in the game, when all the monsters are connected and they make music, electrical energy is transferred from one place to another. The energy travels along the path or circuit.

The presentation provides a series of cognitive supports aligned to the analogy mapping techniques, including presenting visuals of the analogy source and target in a manner that highlights the alignment between the game mechanics and the analogy target (i.e., the game mechanics are next to a picture of the corresponding energy process); also, the notes for teachers embedded in the PowerPoint presentation prompt them to keep the presentation images visible to students as they discuss them, and to gesture between the source and the target as they compare the two images and processes.

Teachers in the comparison condition were guided to provide similar levels of instruction and monitoring and with the same timing as teachers in the comparison during the two class periods after the game and after instruction. Specifically, comparison teachers also conducted group or whole class activities to scaffold student learning and thinking, but focused on activating student prior knowledge about the target concept of focus, rather than student game-related knowledge. In the comparison condition, rather than debriefing about the game (as in the treatment condition), teachers had students describe in words, examples, and pictures their prior knowledge about the science topic they were going to study. Then they engaged in whole-class or small-group discussions about each sequence topic (photosynthesis, heat transfer, or electricity), using formative assessment prompts published by the National Science Teachers Association (NSTA) (Keely, Eberle, & Dorsey, 2008). The student groups had to select and explain their answer to the prompt. For example, when exploring electricity and energy transfer, the prompt asks students to resolve a disagreement between two students who are building a circuit to light a lightbulb: they notice the wires they are using are tangled up in knots. One student thinks they need to untangle the wires because the knots stop the flow of electricity, while the second student thinks the knots will not stop the flow of electricity. Students had to choose an answer and provide a rationale for their answer.

Both conditions also revisited their prior knowledge with teacher prompting and scaffolding after receiving instruction on the target science concepts. After topic instruction and a topic-related activity, instead of the analogy mapping done in the treatment condition, the comparison group engaged in a follow-up discussion moderated and guided by the teacher in which they reviewed their original responses to the science prompt, discussed in their small groups whether they wanted to change their answer based on what they had learned, and explained to the whole class their reason for changing or staying with their original answer.

Before the intervention, we provided a full-day professional development session to both the treatment and comparison teachers together, in which they played one of the games and learned about the instructional materials and activities they would be using. Following this introduction, the two conditions separated, and the treatment group learned about the game debrief and analogy mapping materials, while the comparison group learned about the prior knowledge activity and formative assessment prompts. All of the comparison and treatment condition teachers spent the same amount of time in these full-day professional development sessions. We then provided both the treatment and comparison teachers with additional 1-h webinars about the second and third games and instructional materials, prior to their implementation.

2.2. Sample

The study sample consisted of 11 middle-grade science teachers from three low-income urban school districts in the Northeast United States and 231 of their sixth and seventh-grade students. One goal of this study was to develop and test an intervention that would be usable and that would improve learning for underserved students in low-income schools. We focused on teachers in low-income schools in particular in an attempt to understand how to remedy the persistent science achievement gaps between low-income students and their wealthier peers (Morgan, Farkas, Hillemeier & Maczuga, 2016). Therefore, researchers recruited teachers by contacting science recruiters in three northeast cities that served predominantly low-income students. Researchers also contacted charter school networks in these areas. Any teachers in these geographic areas who taught photosynthesis, electricity, and heat transfer as part of their regular sixth- or seventh-grade science class were eligible to participate in the study. The researcher team invited all interested teachers to attend an information session in person or virtually to describe the study and content of the instructional intervention. All teachers who joined the information session were invited to participate in the study; 11 teachers elected to join the study.[1] While all teachers served students in the study's target population—low-income—because enrollment in the study was based on their own initiative, we assume that these teachers are slightly more motivated than the average teacher population for these schools. They also held more years of science teaching than is typical. Table 2 provides a description of the teachers' backgrounds.

Because this intervention is newly developed, the intention of this study was to test the feasibility and promise of the intervention, and therefore, it is small-scale and not fully powered. Results of our power analysis using the PowerUp! tool (Dong & Maynard, 2013) suggest that a sample of 11 teachers with an average of 17 students per classroom (assuming that student-level pretests will explain 50% of the variance of our outcome and an intra-class-correlation at the classroom level of 0.15) will allow us to detect an effect size of 0.72 SD. Although effect sizes tend to be large for an intervention that relies on a narrow focal topic test or mastery test outcome, such as our primary outcome of interest for this study (Lipsey, Puzio, Yun, Herbert, Steinka-Fry, Cole, et al., 2012), this sample size does not provide sufficient power to detect an unbiased estimate of the effect of the intervention—nor does the non-randomized design. Rather, because of the uncertainty embedded in the study design, our estimate of the magnitude of the effect size from this study must be one piece of evidence taken into account along with evidence regarding whether teachers and students were able to use the intervention as prescribed, and in a manner aligned to the study's primary theory of action (Westlund & Stuart, 2017).

We assigned six of these teachers to the treatment and five to the comparison condition, and matched assignment to the two conditions to ensure balance on criteria related to grade level taught, years of experience teaching, school district, school percent of students eligible for free or reduced-price lunch, grade average on state test scores in math and science, and the time of year the teacher intended to implement the intervention. All 11 teachers and 231 of their sixth- and seventh-grade students completed the study. Table 2 presents descriptive statistics for the sample overall by study condition (treatment or comparison). We found no significant differences between teachers in the treatment and comparison groups on any of the measured demographic characteristics, with two exceptions. Students in the treatment group had slightly lower pre-assessment scores (mean difference = .35 SD; $p < .01$) and were less likely to be in seventh grade compared to the comparison group ($p < .01$). We controlled for these initial differences in our model, described below.

2.3. Measures

2.3.1. Treatment

The primary variable of interest is a teacher-level indicator of whether the teacher was assigned to the treatment group (treatment = 1, comparison = 0).

2.3. Outcomes

We adapted multiple-choice assessments from existing sources to assess student understanding of facts and concepts related to the

[1] One teacher attended the information session and decided not to enroll in the study because his curriculum did not align to the content of focus of the games.

Table 2
Characteristics of participating students for treatment group and comparison group.

Characteristic	Treatment[a]	Comparison[b]
Student Characteristics (n = 231)		
Pre-Assessment Scores***M(SD)*	0.17 (0.97)	−0.19 (1.0)
Gender *(%)*		
Female	47	51
Male	53	49
Grade (%)**		
6th	73	55
7th	27	45
Teacher Characteristics (n = 11)		
Grade *(%)*		
6th	67	60
7th	33	40
District *(%)*		
Large Urban	66	40
Medium Urban 1	17	20
Medium Urban 2	17	40
Class Average Pre-Assessment Scores **M(SD)*	0.20 (0.32)	−0.23 (0.24)
UTOP Summary Score*M(SD)*	3.50 (0.55)	3.25 (0.61)
Average Years Teaching*M(SD)*	14.83 (9.19)	14.60 (4.93)
Average Years Teaching Science*M(SD)*	11.00 (9.23)	9.40 (4.28)
Grade Average State Math Score*M(SD)*	32.83 (12.49)	40.80 (12.29)
Grade Average State ELA Score*M(SD)*	30.00 (15.62)	31.40 (12.18)
School Percent Free and Reduced-Price Lunch*M(SD)*	70.50 (24.13)	74.40 (12.72)
School Percent English Language Learners[c]*M(SD)*	12.33 (15.73)	24.50 (16.84)

Notes.
*p < .05; **p < .01; ***p < .001. Pre-Assessment scores are standardized scores (z-scores).
[a] Teacher n = 6, student n = 118.
[b] teacher n = 5, student n = 113.
[c] missing data from one comparison classroom whose district does not report English language learners.

three topic areas (photosynthesis, heat transfer, and electricity) and to the larger crosscutting energy transfer concept. These online assessments were composed of items from validated, nationally-normed assessments that measure student thinking related to these three concepts of focus, including the Misconceptions-oriented Standards-based Assessment for Teachers (MOSART) concept inventory (Haladyna, 2004; Sadler, 1998; Sadler, Sonnert, Coyle, Cook-Smith, & Miller., 2013) and items from the American Association for the Advancement of Science (AAAS) Project 2061 Science Assessment website (Herrmann Abell & DeBoer, 2011; Stern & Ahlgren, 2002; Stern & Roseman, 2006). We relied on content-based assessments in a traditional multiple-choice format because these kinds of tests are typical of the high-stakes assessments that middle and high school students are required to pass. Moreover, few, if any validated assessments exist that might better capture student thinking processes and understanding of the connections across ideas (Liu, Ryu, Linn, Sato, & Svihla, 2015). We pilot-tested a version of the assessments with all available relevant items from these sources during field tests of the modules in the 2015-16 school year with 85 students in four New York City middle schools. Based on psychometric analysis of the pilot data, we then revised and shortened the assessments: the number of items on the final assessments vary by assessment topic from 16 to 23 questions. Reliability analyses of the assessments indicated sufficient internal reliability (Bacon, 2004; Nunnally, 1978), as measured by Cronbach's alpha (photosynthesis assessment: 0.81; heat transfer assessment: 0.80; electricity assessment: 0.76; energy transfer assessment: 0.75). Teachers administered the energy transfer assessment prior to teaching the three topics, and then administered the corresponding topic-area assessments after teaching each of the three topics. At the end of the three units, teachers administered the same energy assessment as a post-test. We summed the number of items each student answered correctly, and standardized into z-scores to allow for comparisons across assessments.

2.3.3. *Student characteristics*

Student characteristics include a dummy-coded measure of student gender, as reported by the teacher (1 = female, 0 = male), and grade (1 = sixth grade, 0 = seventh grade).

2.3.4. *Teacher characteristics*

We collected data on teachers' background characteristics using a survey that all participating teachers completed prior to participating in the study, which asked about years of experience teaching and years of experience teaching science. We also included a measure of instructional quality prior to implementation of the study, based on the UTeach Observation Protocol (UTOP) for Mathematics and Science, a classroom observation measure with established reliability and validity (Kane & Staiger, 2012). Developed by researchers at the University of Texas, Austin, the UTOP is an observation protocol used to assess the quality of math and science instruction, and is composed of 30 items on a 5-point Likert Scale related to four dimensions: classroom environment, lesson structure, implementation, and science content (Walkington et al., 2012). Researchers established inter-rater reliability of 80% for each code. We created an overall UTOP score for each teacher by computing the average rating across teacher scores on each of the 30 indicators, and because the scores were not normally distributed, and to allow for non-linear relationships, we created a three-level measure of teacher quality, coding teachers in the top and bottom 30th percentiles as “high” and “low” UTOP scores.

2.3.5. Classroom and school characteristics

Classroom characteristics include a continuous measure of classroom average standardized (z-scores) energy pre-test scores and an indicator of whether the school was located in a large or medium urban district (large = 1, medium = 0).

2.4. Analyses

Our analyses focused on examining the differences between the two study conditions in student learning, as measured by multiple choice assessments. We explored the relationship between assignment to the treatment and outcomes on four assessments that measure knowledge related to energy transfer, heat transfer, electricity, and photosynthesis. We analyzed the relationship between student learning and analogy mapping using two-level hierarchical linear models to account for the nested structure of the data, in which students are nested within classrooms/teachers (Raudenbush & Bryk, 2002). We fit these models by maximum likelihood using *xtmixed* in Stata/IC version 15. The analysis models include a treatment effect—whether in the treatment condition—which varies randomly by classroom/teacher, and a random effect for classroom/teacher. In order to control for differences in background characteristics between the comparison and treatment teachers and classrooms, and in the interest of model parsimony given the small sample size, our analyses also incorporate those student and teacher background characteristics most likely to be strong predictors of student learning, including student pre-test scores and gender, whether located in a large urban district or medium one, and grade. We also include treatment pre-test interaction terms in our models to examine whether the relationship between the intervention and student learning depends on students' prior science knowledge—for example, whether low-performing students learned more (or less) from the intervention than did high-performing students.

Our Level-1 model is

$$\text{Assessment}_{\text{ij}} = \beta_{0\text{j}} + \beta_{1\text{j}}(\text{Pretest}_{\text{ij}}) + \beta_{2\text{j}}(\text{Female}_{\text{ij}}) + \varepsilon_{\text{ij}}$$

where *Assessment*$_{ij}$ is the score on the energy, photosynthesis, heat, or electricity assessment for each student *i* of teacher *j* at the end of the corresponding topic area unit. The remaining variables in the Level 1 model are covariates we included in order to statistically adjust for preexisting differences in students, each of which are group-mean centered.

Because this is a teacher-level intervention and the study assigns teachers to the treatment and comparison groups, the test of whether the intervention had an impact on students' assessment scores is specified in the Level 2 (teacher-level) model:

$$\beta_{0\text{j}} = \gamma_{00} + \gamma_{01}(\mathit{Treatment}_j) + \gamma_{02}(\mathit{Sixth\ grade}_j) + \gamma_{03}(\mathit{Large\ Urban}_j) + \text{u}_{0j}$$

$$\beta_{1\text{j}} = \gamma_{10} + \gamma_{11}(\mathit{Treatment}_j)$$

$$\beta_{2\text{j}} = \gamma_{20}$$

$$\beta_{3\text{j}} = \gamma_{30}$$

where *Treatment*$_j$ indicates whether a teacher was in the treatment group and γ01 captures the difference in average assessment scores for treatment and comparison group classrooms.

2.4.1. Missing data

Because the study was conducted over the course of a few months, and teachers administered assessments at four points over the course of these months, students in our sample are missing data on particular measures, and it is likely that these data are not missing completely at random. For example, low-income students are more likely to be absent, and thus more likely to be missing assessment scores. Table 3 displays comparisons of the characteristics of the sample that includes all cases and the sample that includes only the cases that have no missing data. Missing data rates for assessment scores vary from 10 to 15% of cases. While the sample of complete cases is similar to the sample with missing cases, our results suggest that students who are not missing data are more advantaged than are those who are missing data. Students in the complete-case sample, with no missing data, have very slightly higher pre-test and post-test scores on average, compared to the full sample, suggesting students are not missing assessments at random, and thus bias our estimates (Allison, 2002). In other words, our sample is skewed slightly toward higher-achieving students—for example, if the impact of the analogy mapping is larger for higher-achieving students than for other students, this will likely pull the estimated impact upwards. As a sensitivity check, we employed multiple imputation in which all of the missing values are predicted using the existing values for other variables (Enders, 2010; Rubin, 1987) using the “MI” package available for Stata (2017) to generate five imputed datasets (Honaker, King, & Blackwell, 2011). In our imputation models, we included all student-level measures, including pre- and post-assessment scores, student gender, and dummy indicators of the teachers.

Table 3
Comparison of means for full sample and complete case sample.

	Percent of cases missing the measure	Full Sample	Sample with Complete Cases
Study condition %			
Treatment		51.1	55.0
Comparison		48.9	45.0
Gender *(%)*	1.3		
Female		48.7	50.4
Male		51.3	49.6
Grade (%)			
6th		64.1	58.0
7th		35.9	42.0
Energy Pre-Assessment Score [a] *M(SD)*	10.0	0.00 (1.00)	0.02 (0.90)
Energy Post-Assessment Score [a] *M(SD)*	14.7	0.00 (1.38)	0.04 (1.36)
Heat Post-Assessment Score [a] *M(SD)*	10.0	0.00 (1.11)	0.03 (1.11)
Electricity Post-Assessment Score [a] *M(SD)*	11.7	0.00 (1.04)	0.07 (1.04)
Photosynthesis Post-Assessment Score [a] *M(SD)*	11.07	0.00 (1.04)	0.15 (0.99)

[a] Measure is standardized (z-scores).

3. Findings

3.1. Implementation

Our classroom observations (which were videotaped) and teacher reports indicate that teachers followed the three instructional sequences, in both conditions, with fidelity. Based on researcher observations, all students in both conditions played the games for the time specified, and all teachers in both conditions implemented the prior knowledge or game debrief activities and the classroom discussion connecting these discussions to the target concept—including, in the case of the treatment condition, the analogy mapping. Teacher reports during interviews with researchers also suggest that all teachers implemented the additional elements of the instructional sequences, including the hands-on activities and the teacher-led presentation on the focal topic. We also examined teacher use of analogies, in both conditions, during the class period when treatment teachers used the analogy mapping materials and the class period when comparison teachers used the science prompt materials, which we report on elsewhere (Martin, Silander, & Rutter, under review). Our results suggest that teachers and students in the treatment group made six times as many references to analogies during instruction as did students and teachers in the comparison group. Treatment teachers made references to analogies a mean of 40.8 times across the three analogy mapping classes, compared to a mean of 7.0 times in comparison classrooms ($p < .01$). Most (72.2%) of the treatment teachers' analogy references focused on the digital games and their relationship with the target concepts. In contrast, the comparison teachers and students more commonly used other sources for their analogies, mostly from outside the domain of science, while just 22.9% of their references to analogies related to the digital games.

3.2. Results

Table 4 presents our findings regarding the relationship between analogy mapping training and materials and student learning on the four multiple-choice assessments. All estimates are fully adjusted for the student-level covariates described above. Our results suggest that students in the treatment group learned more than did those in the comparison group, as evidenced by multiple-choice assessments. Specifically, students in the treatment group learned more energy and electricity concepts. Moreover, our findings suggest that treatment students' learning about energy concepts broadly varied by students' pre-test scores, such that treatment students with higher pre-test scores learned more than did treatment students with lower pre-test scores (main effect = .72, p < .05; interaction effect = 0.38, p < .05). Treatment students learned more electricity concepts, as measured by the electricity assessment (ES = 0.75, p < .001), but this learning did not appear to vary by pre-assessment scores. Although both of these effect sizes are quite large, the relatively large standard errors in these models (ranging from 0.34 to 0.17 for the main effects) suggest quite a bit of uncertainty in the magnitude of the effect size. Student assessment scores did not vary as a function of our measures of teacher quality (p > .05). We did not find significant differences between the comparison and treatment groups on our measures of learning related to heat transfer (p > .05) and photosynthesis (p > .05) specifically (although the positive coefficients for these models are perhaps suggestive of a similar positive trend).

3.2.1. Sensitivity analysis

It is possible, because students missing outcome data had lower pre-assessment scores on average, and our complete case analysis included a more advantaged population, that our findings are biased toward zero. To explore this issue further, we imputed the missing data and examined the same models using the imputed datasets. Our analyses using these imputed datasets indicate that the estimates of outcomes for cases missing data versus those not missing data are substantively the same.

Table 4
Relationship between treatment and student assessment outcomes.

	Model 1 (energy post-)	Model 2 (heat post-)	Model 3 (electricity post-)	Model 4 (photosyn-thesis post-)
Student level: Within-class effects				
Pre-test[a]	0.38***	0.32**	0.43***	0.47***
	(0.11)	(0.11)	(0.17)	(0.09)
Teacher/class level: Intervention				
Treatment	0.72*	0.30	0.75***	0.29
	(0.34)	(0.20)	(0.17)	(0.19)
Pre-test[a]* treatment	0.38*	0.27	0.02	0.00
	(0.11)	(0.15)	(0.14)	(0.13)
Sixth grade[b]	−1.46***	−0.67***	−0.57***	−0.94***
	(0.34)	(0.20)	(0.16)	(0.19)
Urban[c]	−0.74*	0.07	−0.21	−0.14
	(0.34)	(0.19)	(0.16)	(0.19)
Intercept	−0.35	−0.17	−0.39***	−0.11
	(0.25)	(0.14)	(0.12)	(0.14)
Random effects				
Teacher	0.23	0.04	0.02	0.04
(se)	(0.13)	(0.04)	(0.03)	(0.04)
Residual	0.88	0.93	0.78	0.67
(se)	(0.10)	(0.10)	(0.09)	(0.07)
Student n	177	189	182	183
Teacher n	11	11	11	11

*p < .05; **p < .01; ***p < .001.

[a] Assessment scores are standardized scores (z-scores).

[b] Compared to seventh grade.

[c] Compared to small city.

4. Discussion

Researchers who have studied the use of digital games in education have found that games can be effective tools for supporting student learning (Clark et al., 2016), but researchers who focus on the integration of digital games in real-world settings have observed that teachers need strategies for connecting digital games to disciplinary content and concepts to support learning (Bell & Gresalfi, 2017a; Rowe et al., 2016). Certainly, digital games can provide engaging experiences for students, but positive learning outcomes rely on good instruction, and good instruction is based on an understanding of how people learn (National Research Council, 1999). To inform their designs, game developers, like developers of any other instructional materials, can look to the research that cognitive and developmental psychologists and learning scientists have been doing for decades to build theories identifying key factors associated with effective instructional approaches (Reese et al., 2015). In this case, because the games in the study were designed to have features analogous to science concepts, we looked to the research on analogies (Gentner & Maravilla, 2018; Richland & Begolli, 2016; Richland et al., 2007) to inform our study. Not only did these provide theoretical grounding for this study, but specific instructional approaches we sought to test.

Though many studies of digital games in education focus more on game design factors than teachers' instructional practice, studies that have looked at teachers' practice found positive student outcomes when teachers use “bridging activities,” in which teachers have explicit discussions about how games are related to instructional content (Rowe et al., 2016), and when teachers have a deep understanding of the digital game they use (Bell & Greselfi, 2017b). Similarly, our study found that students whose teachers used analogy mapping techniques and related materials to connect digital games to disciplinary content performed better on some assessments of science concepts than students who played the same games but whose teachers did not do the analogy mapping. The findings from this study suggest that using these instructional techniques holds promise as an approach to integrating digital games into instruction. As noted above, we found that the professional development and instructional materials resulted in the treatment group making references to analogies six times more often than did the comparison group (Martin, Silander, & Rutter, under review). Based on our understanding of the literature, we theorized that, when students had a proper understanding of the game and then experienced the analogy mapping instruction, they would be able map visualizations from the game to the science concepts and transfer the game knowledge to the new domain. Our positive findings about increased student learning on some assessments suggest that this might be the case, especially since student performance did not vary by teacher quality, or by teacher experience.

It is interesting that the positive impact was found for the electricity game, but not for the heat transfer or photosynthesis games. It is not clear why this is the case. Each teacher in the treatment and comparison groups used similar instructional methods for each of the games. Not only did they have a consistent set of materials (games, instructional PowerPoint, and in the case of the treatment group, debrief and analogy mapping PowerPoints) but they integrated those into their particular teaching style, which was consistent across each sequence. It is possible that a concept such as how energy moves through an electrical circuit is particularly suited to being taught with a visualization. It may be that the treatment teachers felt more confident in their understanding of current electricity and could therefore elaborate on the visualization more fully than the other topics. It could also be that the electricity game, *Monster Music*, was better designed than the other two games, or has an analogy that was more illuminating for students.

In order to understand in more detail the relationship between the instructional scaffolding and students' abilities to make the game analogies, we assessed student understanding of the central game mechanisms after the game debrief (treatment) or prior knowledge activity (comparison) using an exit ticket that asked how the game related to the science topic. Our findings from these exit tickets suggest that debriefing the game helped students better understand the photosynthesis and electricity games but not heat transfer. Specifically, following the treatment group's debrief of the game, students in the treatment group were more likely to be able to describe how the games related to electricity and photosynthesis compared to the comparison group (ES = .52 and .43 SD, respectively, $p < .001$). These findings suggest that at least for some games, ensuring that students have an understanding of the source analog is an important instructional scaffold. However, in contrast, we found no difference between the two groups in their understanding of the heat transfer analogy following the debrief; we hypothesize that this difference is because the heat transfer game analogy was the least distal analogy and that the important mechanisms were easiest to identify. Thus, it seems possible that game design could be a factor when considering the amount of support students need to make the analogies—students need more support to understand the source analogy with games that offer more distal analogies.

The finding that the game with the least closely aligned analogy showed the most positive impact is consistent with the research. Genter, Bowdle, Wolff, and Boronat (2001) found that analogies tend to be more effective when the source and target come from very different domains, because they demand greater conceptual effort. Making an analogy between monsters holding hands to create music and electrical current required quite a conceptual leap, while the actions in the heat transfer game (an intergalactic zookeeper warming eggs to the right temperature to hatch) and the photosynthesis game (a robot making glucose to provide energy to travel through a cave) were more closely aligned to those concepts. However, the positive findings on the broader energy transfer assessment suggest that the students in the treatment condition gained some level of understanding across the board about the larger concept underlying these three forms of energy transfer, suggesting that using these techniques with digital games is an effective method for connecting game-world learning to real-world conceptual understanding. The fact that student learning from the analogy mapping techniques appeared to vary on the energy assessments such that students with more baseline knowledge learned more from analogy mapping than student with lower levels of initial understanding suggests that there may also be a baseline amount of knowledge that students need to have in order to learn from a game. (The interaction between prior knowledge and learning might also explain why sixth grade students learned less than seventh grade students.) This interaction between prior knowledge and learning was not evident for electricity content, however, and this lack of consistency across the assessments suggests that more research is needed into the relationship between prior knowledge and student ability to learn from digital science games.

The limitations of the study, particularly the small sample size, self-selection into the study and lack of randomized assignment, prevent us from making any causal claims about these techniques. Rather, we can only say that the results indicate that using techniques such as these shows promise as scaffolding for student learning with digital games. Because the study was logistically challenging, involving videotaping of classroom instruction in three different cities and many other days of data collection, we could not have had a larger sample at this stage of the research. However, the intervention itself is not complex (Martin, Pasquale, & Silander, 2018), and could in future be scaled to a larger number of classrooms for a more comprehensive efficacy study. A second limitation of the study relates to our inability to generalize to a broad population of teachers because of the small sample size and the fact that teachers self-selected into the study.

Despite the limitations, however, the findings from this study have broad implications for researchers, educators, and game developers wishing to understand how to integrate digital games effectively into classroom instruction and how to help students in this age group learn challenging science concepts. Analogies can enable students to gain an understanding of new, abstract concepts, but they are most likely to improve learning when the comparative relationships between the source and target are made explicit and presented visually and repeatedly. Digital games can serve as compelling, shared visual, interactive experiences that can be the foundation of an engaging analogy. However, student engagement in and mastery of a game on its own will not necessarily lead to knowledge transfer. That likely requires very deliberate, scaffolded discussions, supported by visual materials, about how the game maps to the analogous concepts.

Funding

This work was supported by the National Science Foundation (Award # DRL-1252382) and the U.S. Department of Education's Institute for Education Sciences (Award # R305C080022).

References

Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology, 55*(1), 193–196.

Andersson, B. (1990). Pupils' conceptions of matter and its transformations (age 12–16). *Studies in Science Education, 18*, 53–85.

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *The Quarterly Journal of Experimental Psychology, 70*(10), 2007–2025. https://doi.org/10.1080/17470218.2016.1219752.

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research Evaluation, 9*(3), Accessed 10.10.2016 http://pareonline.net/getvn.asp?v = 9&n = 3.

Barzalai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education, 70*(1), 65–79.

Bell, A., & Gresalfi, M. (2017a). The role of digital games in a classroom ecology: Exploring instruction with video games. In M. F. Young, & S. T. Slota (Eds.). *Exploding the Castle: Rethinking how video games and game mechanics can shape the future of education* (pp. 67–92). Charlotte, NC: Information Age, Inc.

Bell, A., & Greselfi, M. (2017b). Experience with videogames: How experience impacts classroom integration. *Technology, Knowledge and Learning, 22*(3), 513–526.

Braasch, J. L. G., & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science texts. *Discourse Processes, 47*, 447–479. https://doi.org/10.1080/01638530903420960.

National Research Council (1999). How people learn: Brain, mind, experience, and school. In J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). *Committee on developments in the science of learning.* Washington, DC: National Academy Press.

Callies, S., Gravel, M., Beaudry, E., & Basque, J. (2017). Logs analysis of adapted pedagogical scenarios generated by a simulation serious game architecture. *International Journal of Game-based Learning, 7*(2), 1–19. https://doi.org/10.4018/IJGBL.2017040101.

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.). *International handbook of research on conceptual change* (pp. 61–82). New York, NY: Routledge.

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

Cooperrider, K., Gentner, D., & Goldin-Meadows, S. (2017). Analogical gestures foster understanding of causal systems. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *Proceedings of the 39th annual Meeting of the cognitive science society* (pp. 240–245). Austin, TX: Cognitive Science Society.

Culp, K. M., Martin, W., Clements, M., & Presser, A. L. (2015). Testing the impact of a pre-instructional digital game on middle-grade students' understanding of photosynthesis technology. *Knowledge and Learning, 20*(1), 5–26.

DeWolf, M., Son, J. Y., Bassok, M., & Holyoak, K. J. (2017). Relational priming based on a multiplicative schema for whole numbers and fractions. *Cognitive Science, 41*, 2053–2088.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Gee, J. P. (2013). *Good video games and good learning: Collected essays on video games, learning and literacy.* New York: Peter Lang.

Gee, J. P. (2014). *What video games have to teach us about learning and literacy* (2nd ed.). New York, NY: St. Martin's Press.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.). *The analogical mind: Perspectives from cognitive science* (pp. 199–253). Cambridge, MA: MIT Press.

Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball, & V. A. Thompson (Eds.). *International handbook of thinking & reasoning* (pp. 186–203). New York, NY: Psychology Press.

Gentner, D., & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.). *Encyclopedia of human behavior* (pp. 130–136). (2nd ed.). London, UK: Academic Press.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences, 20*(2), 169–206. https://doi.org/10.1080/10508406.2010.508029.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hattie, J. A. C., & Yates, G. R. (2013). *Visible learning and the science of how we learn.* Oxford, UK: Routledge.

Hayes, E. R., & Gee, J. P. (2012). Passionate affinity groups. In C. Steinkuhler, K. Squire, & S. A. Barab (Eds.). *Games, learning and society* (pp. 129–153). Cambridge, MA: Cambridge University Press.

Herrmann Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemical Education Research and Practice, 12*, 184–192.

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1–47.

Jaeger, A. J., Taylor, A. R., & Wiley, J. (2016). When, and for whom, analogies help: The role of spatial skills and interleaved presentation. *Journal of Educational Psychology, 108*(8), https://doi.org/10.1037/edu0000121 1211–1139.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching.* Seattle, WA: Bill and Melinda Gates Foundation.

Keely, P., Eberle, F., & Dorsey, C. (2008). Uncovering student ideas in science. *Another 25 formative assessment probes. ume 3*. Arlington, VA: NSTA Press.

Law, V., & Chen, C.-H. (2016). Promoting science learning in game-based learning with question prompts and feedback. *Computers & Education, 103*, 134–143.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000).* Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Liu, O. L., Ryu, K., Linn, M. C., Sato, E., & Svihla, E. (2015). Measuring knowledge integration learning of energy topics. *International Journal of Science Education, 37*(7), 1044–1066.

Lu, A. S., Buday, R., Thompson, D., & Baranowski, T. (2016). What type of narrative do children prefer in active video games? An exploratory study of cognitive and emotional responses. In S. Tettegah, & W. D. Huang (Eds.). *Emotions, technology, and digital games* (pp. 137–156). London: Academic Press.

Martin, W., Pasquale, M., & Silander, M. (2018). Mapping digital games to science instruction. *Science Scope, 41*(7), 88–95 National Science Teachers Association.

Martin, W., Silander, M. Culp, K., Brunner, C. & Parris, J. (in press).Designing instructional supports for digital science games: Visualizing and mapping analogies. In M. J. Bishop., E. Boling, J. Elen, V. Svihla (Eds.) Handbook of research on educational communications and technology (5th ed.).

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher, 45*(1), 18–35.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ozay, E., & Oztas, H. (2003). Secondary students' interpretations of photosynthesis and plant nutrition. *Journal of Biological Education, 37*(2), 68–70.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology, 46*(1), 98–122.

Richland, L. E., & Begolli, K. N. (2016). Analogy and higher-order thinking: Learning mathematics as an example. *Policy Insights from the Behavioral and Brain Sciences (PIBBS), 3*(2), 160–168. https://doi.org/10.1177/2372732216629795.

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 177–192.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science, 316*(5828), 1128–1129.

Rowe, E., Asbell-Clarke, J., Bardar, E., Kasman, E., & MacEachern, B. (2014). Crossing the bridge: Connecting game-based implicit science learning to the classroom. *Paper presented at the 10th annual Games + Learning + Society conference in Madison, WI*.

Rowe, E., Bardar, E., Asbell-Clarke, J., Shane-Simpson, C., & Roberts, S. (2016). Building bridges: Teachers leveraging game-based implicit science learning in physics classrooms. In D. Russell, & J. Laffey (Eds.). *Handbook of research on gaming trends in P-12 education*Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-9629-7.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265–296.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle-school physical science classrooms. *American Educational Research Journal, 50*(5), 1020–1049. https://doi.org/10.3102/0002831213477680.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115–163.

Squires, A., Driver, R., & Rushworth, P. (Eds.). (1994). *Making sense of secondary science: Research into children's ideas*. London, UK: Routledge.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching, 39*(9), 889–910.

Stern, L., & Roseman, J. R. (2006). Improving the alignment of curriculum and assessment to national science standards. In D. W. Sunal, E. L. Wright, & Series (Vol. Eds.), *Research in science education: Vol. 2*, (pp. 301–324). Greenwich, CT: Information Age Publishing The impact of state and national standards on K–12 science teaching.

Sun, C.-T., Chen, L.-X., & Chu, H.-M. (2018). Associations among scaffold presentation, reward mechanisms and problem-solving behaviors in game play. *Computers & Education, 119*, 95–111. https://doi.org/10.1016/j.compedu.2018.01.001.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*(1), 4–28.

Travis, R. (2017). What Homeric Epic can teach us about educational affordances of interactive narrative. In M. F. Young, & S. T. Slota (Eds.). *Exploding the Castle: Rethinking how video games and game mechanics can shape the future of education* (pp. 19–37). Charlotte, NC: Information Age, Inc.

Van Der Meij, H., Leemkuil, H., & Li, J.-L. (2013). Does individual or collaborative self-debriefing better enhance learning from games? *Computers in Human Behavior, 29*(6), 2471–2479.

Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom: Insights from cognitive science. *Mind, Brain, and Education, 9*(2), 100–106. https://doi.org/10.1111/mbe.12080.

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., et al. (2012). Development of the UTeach observation protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach preparation program. Retrieved from https://uteach.utexas.edu/sites/default/files/files/2013%20Classroom%20Environment(1).docx.

Westlund, E., & Stuart, E. A. (2017). The nonuse, misuse and proper use of pilot studies in experimental evaluation research. *American Journal of Evaluation, 38*(2), 246–261.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265.

Young, M. F., & Slota, S. T. (2017). Castle upon a hill. In M. F. Young, & S. T. Slota (Eds.). *Exploding the Castle: Rethinking how video games and game mechanics can shape the future of education* (pp. 3–18). Charlotte, NC: Information Age, Inc.

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research, 82*, 61–89. https://doi.org/10.3102/0034654312436980.