Review of Educational Research
March 2016, Vol. 86, No. 1, pp. 79–122
DOI: 10.3102/0034654315582065

Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis

Douglas B. Clark, Emily E. Tanner-Smith, and Stephen S. Killingsworth

Vanderbilt University

In this meta-analysis, we systematically reviewed research on digital games and learning for K–16 students. We synthesized comparisons of game versus nongame conditions (i.e., media comparisons) and comparisons of augmented games versus standard game designs (i.e., value-added comparisons). We used random-effects meta-regression models with robust variance estimates to summarize overall effects and explore potential moderator effects. Results from media comparisons indicated that digital games significantly enhanced student learning relative to nongame conditions ($\bar{g}$ *= 0.33, 95% confidence interval [0.19, 0.48],* k *= 57,* n *= 209). Results from value-added comparisons indicated significant learning benefits associated with augmented game designs (* $\bar{g}$ *= 0.34, 95% confidence interval [0.17, 0.51],* k *= 20,* n *= 40). Moderator analyses demonstrated that effects varied across various game mechanics characteristics, visual and narrative characteristics, and research quality characteristics. Taken together, the results highlight the affordances of games for learning as well as the key role of design beyond medium.*

Keywords: digital games, learning, meta-analysis, systematic review

In 2006, the Federation of American Scientists (FAS) issued a widely publicized report stating that games as a medium offer powerful affordances for education. The report encouraged private and governmental support for expanded research into complex gaming environments for learning. A special issue of *Science* in 2009 echoed and expanded this call (Hines, Jasny, & Mervis, 2009). Studies have demonstrated the potential of digital games to support learning in terms of conceptual understanding (e.g., Barab et al., 2007; Klopfer, Scheintaub, Huang, Wendel, & Roque, 2009), process skills and practices (e.g., Kafai, Quintero, & Feldon, 2010; Steinkuehler & Duncan, 2008), epistemological

understanding (e.g., Squire & Jan, 2007; Squire & Klopfer, 2007), and players' attitudes, identity, and engagement (e.g., Barab et al., 2009; Dieterle, 2009; Ketelhut, 2007). Reports by the National Research Council (NRC) and others (e.g., Honey & Hilton, 2010; Martinez-Garza, Clark, & Nelson, 2013; Young et al., 2012) have acknowledged this potential but also acknowledge the unevenness of systematic evidence for games as learning tools.

In the current meta-analysis, we systematically reviewed research on digital games and learning for K–16 students in light of the recent NRC report on education for life and work in the 21st century (Pellegrino & Hilton, 2012). We synthesized comparisons of game conditions versus nongame conditions (i.e., media comparisons) as well as comparisons of augmented game designs versus equivalent standard game designs (i.e., value-added comparisons). Meta-regression models were used to assess the possible moderating effects of participant characteristics, game condition characteristics, and research quality characteristics.

Alignment With Recent Related Meta-Analyses

The current meta-analysis extends and refines the findings of three recent meta-analyses relevant to the impact of games on learning.[1] We first provide an overview of these three relevant meta-analyses to frame the relationships, contributions, and research questions of the current meta-analysis. The first meta-analysis, by Vogel et al. (2006), synthesized results from 32 studies from 1986 to 2003, focusing on pretest–posttest comparisons of cognitive and attitudinal outcomes in games and simulations for age-groups spanning preschool through adult. Vogel et al. described computer games and simulations as follows:

> A computer game is defined as such by the author, or inferred by the reader because the activity has goals, is interactive, and is rewarding (gives feedback). Interactive simulation activities must interact with the user by offering the options to choose or define parameters of the simulation then observe the newly created sequence rather than simply selecting a prerecorded simulation. (p. 231)

The synthesized studies compared games and simulations to traditional classroom teaching. Moderator variables included gender, learner control, type of activity (game vs. simulation), age, visual realism, and player grouping (individual vs. group).

Overall, Vogel et al. (2006) found that games and simulations led to higher cognitive outcomes ($z = 6.05$) and attitudinal outcomes ($z = 13.74$) than traditional instruction. Although the number of included studies limited other conclusions, Vogel et al.'s findings suggested (a) no differences across age, gender, visual realism, and type of activity but (b) potential differences in terms of learner control and player grouping. In particular, effect sizes were higher for studies involving individual students versus groups, and effect sizes were lower for studies where learners had less control.

The second meta-analysis, by Sitzmann (2011), synthesized results from 65 studies from 1976 to 2009, focusing on pretest–posttest comparisons of self-efficacy, declarative knowledge, procedural knowledge, and retention in simulation

games for adult workforce trainees. Sitzmann defined simulation games as "instruction delivered via personal computer that immerses trainees in a decision-making exercise in an artificial environment in order to learn the consequences of their decisions" (p. 492). Comparison conditions in the synthesized studies ranged from no-training control conditions to alternative instructional method conditions. Theoretical moderator variables included entertainment value, whether the simulation game instruction was active or passive, whether or not trainees had unlimited access to the simulation game, whether the simulation game was the sole instructional method, and whether the instructional methods in the comparison group were active or passive. Methodological moderator variables included random assignment to experimental condition, rigor of the study design, publication status, and year of the publication/presentation.

Sitzmann (2011) found that self-efficacy was significantly higher ($d = 0.52$) as were declarative knowledge ($d = 0.28$), procedural knowledge ($d = 0.37$), and retention ($d = 0.22$) for trainees receiving instruction via a simulation game than for trainees in the comparison conditions. The three cognitive outcomes were found to not differ significantly from one another. In terms of moderators, all the theoretical moderators except entertainment value proved significant. Trainees with simulation games learned more, relative to the comparison group, when (a) simulation games were active rather than passive learning experiences, (b) trainees had unlimited access to the simulation game, and (c) the simulation game was supplemented with other instructional methods. The comparison group learned more than the simulation game group when the comparison group received instruction that actively engaged them in the learning experience. In terms of methodological moderators, only publication status was significant, demonstrating that simulation game groups outperformed comparison groups to a greater extent in published studies than in unpublished studies.

The third meta-analysis, by Wouters, van Nimwegen, van Oostendorp, and van der Spek (2013), analyzed 39 studies from 1990 to 2012 focusing on pretest–posttest and posttest-only comparisons of knowledge, skills, retention, and motivation outcomes in serious games for a wide range of age-groups. Wouters et al. defined serious games as follows:

> We describe computer games in terms of being interactive (Prensky, 2001; Vogel et al., 2006), based on a set of agreed rules and constraints (Garris et al., 2002), and directed toward a clear goal that is often set by a challenge (Malone, 1981). In addition, games constantly provide feedback, either as a score or as changes in the game world, to enable players to monitor their progress toward the goal (Prensky, 2001). . . . In speaking of a serious (computer) game, we mean that the objective of the computer game is not to entertain the player, which would be an added value, but to use the entertaining quality for training, education, health, public policy, and strategic communication objectives (Zyda, 2005). (p. 250)

Comparison conditions in the synthesized studies included conventional instruction methods such as lectures, reading, drill and practice, or hypertext learning environments. Theoretical moderator variables included active versus

passive instruction in comparison groups, presence of additional nongame instruction in game conditions, level of visual realism in game conditions, level of narrative in game conditions, number of training sessions, group size, instructional domain, and age. Methodological moderator variables included publication source, random assignment, and pretest–posttest versus posttest-only assessment.

Wouters et al. (2013) found that serious games were more effective than conventional instruction in terms of learning ($d = 0.29$) and retention ($d = 0.36$) but found no evidence that they were more motivating ($d = 0.26$). Moderator analyses revealed that games were more effective when games were supplemented with other instruction, when multiple training sessions were involved, and when players worked together in groups. In terms of visual realism, schematic serious games were significantly more effective than cartoon-like or realistic games. In terms of narrative, findings were not significant but suggested that serious games without a narrative might be more effective than serious games with a narrative. Interestingly, serious games showed larger gains when compared to mixed instruction than when compared to passive instruction. In terms of methodological moderators, random assignment and publication source attenuated the effect of serious games, but there were no differences in effects for pretest–posttest versus posttest-only assessments.

Core Hypotheses

Drawing on results from these prior meta-analyses, the present meta-analysis sought to extend and refine our understanding of the effects of digital games on learning outcomes for K–16 students. Methodologically, the current meta-analysis expanded on prior work by broadening the scope of the literature surveyed. Research on games for learning spans many fields. We thus selected databases spanning Engineering, Computer Science, Medicine, Natural Sciences, and Social Sciences in an effort to capture this breadth while focusing on research published between 2000 and 2012 in light of the dramatic evolution of digital games for learning over the past decade. Furthermore, the current meta-analysis provides a specific and distinct focus on (a) digital games, (b) K–16 students, and (c) cognitive, intrapersonal, and interpersonal learning outcomes. The current study therefore builds on the prior meta-analyses by expanding the scope of constituent studies while focusing on an overlapping but distinct cross section of the research literature with a tighter focus on games and learning by K–16 students (Table 1). Overall, based on the prior meta-analyses, we predicted that game conditions would be associated with better learning outcomes than nongame conditions in media comparisons (Core Hypothesis 1).

Whereas prior meta-analyses have focused exclusively on comparisons of game conditions versus nongame control conditions—which Mayer (2011), calls media comparisons—the present study focused on value-added comparisons also. Value-added comparisons measure the efficacy of a standard version of a game relative to an enhanced version augmented to test a theoretical design proposition (Mayer, 2011). Wouters et al. (2013) expressed the need for analyses of value-added studies in their discussion. The present study thus moved beyond a sole focus on media comparisons to also assess the contribution of design to learning.

TABLE 1

Characteristics of recent meta-analyses on games for learning: Overlapping but distinct lenses

Authors	Learning environments	Scope	Study count	Years included	Demographics	Data
Vogel et al. (2006)	Computer games and simulations	Media comparison	32	1986–2003	All age-groups	Pretest–posttest cognitive and attitudinal
Sitzmann (2011)	Simulation games	Media comparison	65	1976–2009	Adult workforce trainees	Pretest–posttest, posttest-only self-efficacy, declarative knowledge, procedural knowledge, retention
Wouters, van Nimwegen, van Oostendorp, and van der Spek (2013)	Serious games	Media comparison	39	1990–2012	All age-groups	Pretest–posttest, posttest-only knowledge, retention, and motivation
Present study	Digital games	Media comparison and value-added	69	2000–2012	K–16 students	Pretest–post cognitive, intrapersonal, and interpersonal

Although it might appear common sense that versions of a game that have been augmented to support learning should outperform standard versions of those games, the role of design has often been de-emphasized in debates over whether digital games are *better* or *worse* than traditional instruction (as highlighted by the preponderance of media comparisons relative to value-added comparisons available in the literature on games for learning). We analyzed value-added comparisons in the current meta-analysis specifically to emphasize the importance of a shift toward more focused comparisons of game designs. We predicted that conditions involving theoretically augmented game designs for learning would be associated with better learning outcomes than conditions involving standard versions of those games in value-added comparisons (Core Hypothesis 2).

Beyond these two core hypotheses, the present study analyzed the potential moderating effects of (a) general study characteristics, (b) game mechanics characteristics, (c) visual and narrative characteristics, and (d) research quality characteristics. These moderator analyses explored the relationships between design features and learning outcomes. The number of media comparisons that met the eligibility criteria (outlined in the Method section) was sufficient to support moderator analyses of general study characteristics, game mechanics characteristics, and visual and narrative characteristics. The number of value-added and media comparisons that met eligibility criteria was sufficient to support moderator analyses in terms of research quality characteristics. We elaborate on the moderator analyses and hypotheses in the following sections.

Moderator Analyses of General Study Characteristics

The present meta-analysis examined three general study characteristics as potential moderators of the effects of digital games on learning. Specifically, we examined game duration, presence of nongame instruction in game conditions, and player grouping. All these moderators were identified from prior meta-analyses on this topic.

With regard to duration of game play, Sitzmann (2011) found that media comparisons in which trainees had unlimited access to the game demonstrated significantly better learning outcomes than media comparisons in which the trainee had limited access to the game. Similarly, Wouters et al. (2013) found that (a) game conditions where participants interacted with the game for more than one session demonstrated significantly better outcomes relative to the nongame control conditions, but (b) game conditions where participants engaged with the game for only one session did not demonstrate significantly better outcomes relative to the nongame control conditions.

Whereas the Sitzmann (2011) comparisons emphasized additional time on task and increased learner control relative to the comparison groups, Wouters et al. (2013) focused on a combination of spaced versus massed learning (cf. McDaniel, Fadler, & Pashler, 2013) and the potential for greater incremental value of additional time in games compared to the incremental value of additional time in associated control conditions. As Wouters et al. (2013) explained, "It is plausible that, in comparison to that of conventional instruction methods, the effectiveness of serious games in terms of learning pays off only after multiple training sessions in which the players get used to the game" (p. 251). The studies synthesized in the

current analysis involve primarily equivalent amounts of total time in experimental and control conditions, and thus our analyses align more closely with the relationship between experimental and control conditions in the Wouters et al. (2013) analyses. Based on these findings, we predicted that game conditions involving increased duration and number of game play sessions would be associated with better learning outcomes in media comparisons (Moderator Hypothesis 1a).[2]

In terms of supplemental nongame instruction, two prior meta-analyses (Sitzmann, 2011; Wouters et al., 2013) found that comparisons where game conditions included supplemental nongame instruction demonstrated better learning outcomes (relative to nongame conditions) than comparisons where the game conditions did not include nongame instruction. Given the importance of verbalization for learning (Wouters, Paas, & van Merriënboer, 2008), and the effects of supplemental instruction on learning observed in prior meta-analyses, we predicted that game conditions that include nongame instruction would be associated with better learning outcomes than game conditions that do not include nongame instruction in media comparisons (Moderator Hypothesis 1b).

In terms of player group structures in game conditions, Vogel et al. (2006) found significant learning outcomes for single-player as well as for collaborative conditions relative to nongame conditions and reported a trend toward larger effect sizes with solitary players but did not report analyses comparing effect size magnitudes between the two player grouping structures. Based on this trend, and given the ambiguity in prior research on the benefits of collaborative play (e.g., Schwartz, 1995; van der Meij, Albers, & Leemkuil, 2011), Wouters et al. (2013) hypothesized that single-user play would outperform group play but found that learners who played serious games in a group learned more than learners who played alone. In the current meta-analysis, we therefore predicted that collaborative game conditions would be associated with better learning outcomes than single-player game conditions in media comparisons (Moderator Hypothesis 1c).

Moderator Analyses of Game Mechanics Characteristics

In addition to exploring general study characteristics, we explored game design mechanics as potential moderators of game effects on learning outcomes. Specifically, we explored broad sophistication of game mechanics (simple gamification of academic tasks vs. more elaborate game mechanics), variety of player actions (focused games like *Tetris* vs. games like *SimCity* where players engage in a wider variety of actions), intrinsic/extrinsic design properties, and degree of scaffolding. We predicted that game conditions involving increased sophistication of game mechanics, variety of player actions, intrinsic integration of the game mechanic and learning mechanic, and specific/detailed scaffolding would be associated with better learning outcomes in media comparisons (Moderator Hypotheses 2a–2d).

Moderator Analyses of Visual and Narrative Game Characteristics

Results from prior meta-analyses examining the effects of digital games on learning have yielded inconsistent and conflicting findings regarding the moderating effect of visual realism. We coded three unique visual characteristics: visual realism, camera perspective, and anthropomorphism. The relevance of camera

viewpoint for learning was included because of the numerous reports that have shown that individuals who play first-person perspective "shooter" games, but not other games, demonstrate improvement on certain visual cognitive tasks (e.g., Feng, Spence, & Pratt, 2007; Green & Bavelier, 2006, 2007). Anthropomorphism was included because of numerous findings suggesting that anthropomorphic attributes affect a range of perceptual, cognitive, and social tasks (e.g., Heider & Simmel, 1944; Killingsworth, Levin, & Saylor, 2011; Mahajan & Woodward, 2009).

In addition to including these visual characteristics, we examined the narrative characteristics of each game condition. Overarching research on learning has supported the inclusion of narrative context in the sense of situating and anchoring learning in context (e.g., Bransford, Brown, & Cocking, 2000; Bransford, Sherwood, Hasselbring, Kinzer, & Williams, 1990; Brown, Collins, & Duguid, 1989). Furthermore, the role of narrative in games for learning remains a central focus of the field (e.g., Dickey, 2006; Echeverria, Barrios, Nussbaum, Amestica, & Leclerc, 2012; Lim, 2008; Malone & Lepper, 1987).

Based on the findings of Wouters et al. (2013), however, we predicted that game conditions involving increased visual realism, anthropomorphism, camera perspective, story relevance, and story depth would be associated with smaller learning outcomes in media comparisons (Moderator Hypotheses 3a–3e). In addition, we predicted that game conditions involving increased overall contextualization would be associated with smaller learning outcomes in media comparisons (Moderator Hypothesis 3f).

Research Characteristics in Value-Added and Media Comparisons

Beyond study and game characteristics, we also explored whether research quality was associated with better or smaller effects in the media comparisons and value-added comparisons. Prior meta-analyses have noted issues with the methodological quality of the primary studies in the games literature (Vogel et al., 2006) and have noted that the beneficial effects of serious games may be attenuated in studies with random assignment versus quasi-experimental designs (Wouters et al., 2013). Based on prior findings with research characteristics, we predicted that comparison condition quality, sufficient condition reporting, sufficient reporting of methods and analyses, overalignment of assessment with game, assessment type, and study design will be associated with learning outcomes in value-added and media comparisons (Moderator Hypotheses 4a–4f).

In-Depth Exploration of Variability in the Effects of Games on Learning

To test the hypotheses outlined above, we explored variability in the effects of digital games on learning outcomes by employing a recently developed statistical technique for robust variance estimation (RVE) in meta-regression (Hedges, Tipton, & Johnson, 2010; Tipton, 2013). This technique permits the inclusion of multiple effect sizes from the same study sample within any given meta-analysis—a common occurrence in meta-analyses in the educational and social sciences (e.g., Tanner-Smith & Tipton, 2014; Tanner-Smith, Wilson, & Lipsey, 2013; Wilson, Tanner-Smith, Lipsey, Steinka-Fry, & Morrison, 2011). This approach avoids loss of information associated with dropping effect sizes (to ensure their statistical independence) and does not require information about the covariance structure of

effect size estimates that would be necessary for the use of multivariate meta-analysis techniques (see Tanner-Smith & Tipton, 2014, for a discussion).

Method

Inclusion and Exclusion Criteria

Digital Game

Eligible studies were required to include at least one comparison of a digital game versus nongame condition or at least one comparison of an augmented game design versus equivalent standard game design (but these two types of comparisons were always analyzed separately). Studies were required to designate explicitly the environment as a *game*, and the term *game* or *games* needed to appear in the abstract or title of the report. A digital game was defined for the purposes of the present meta-analysis as a digital experience in which the participants (a) strive to achieve a set of fictive goals within the constraints of a set of rules that are enforced by the software, (b) receive feedback toward the completion of these goals (e.g., score, progress, advancement, win condition, narrative resolution), and (c) are intended to find some recreational value.

Hybrid augmented reality games that used digital platforms to create games in physical space were eligible, but physical games with no digital platform were excluded (e.g., board games). Interventions that focused primarily on teaching youth to create or program games were not included for the present analyses because these approaches were considered distinct (and potentially more powerful) in light of their closer alignment with design-based learning (e.g., Kafai, 2006). In terms of recreational value, we do not imply joviality—games, like books and movies, can be serious or sad, thus communicating a powerful experience and message while doing so in a way that draws in people willingly to play for the sake of play (cf. Young et al., 2012). In terms of simulations, while most games have some form of simulation within them, the current meta-analysis includes only studies where (a) the digital environment in the study meets the eligibility criteria definition of a game outlined above and (b) the digital environment is explicitly referred to by the authors of that study as a game in the title or abstract. Thus, simulations that do not meet the game eligibility criteria outlined above are not included in the current meta-analysis.

Participants

Eligible participant samples included students in K–16, ages 6 to 25. Participants had to be students in a K–12 institution or enrolled in postsecondary school. Studies of participants beyond the K–16 grade range were not eligible. Studies focusing on samples from specific clinical populations of students (e.g., autism spectrum) were also excluded.

Research Designs

Because the focus of the meta-analysis was on making causal inferences regarding the effects of digital games on learning, only those studies using randomized controlled trial and controlled quasi-experimental research designs were eligible for inclusion.

Learning Outcomes

Eligible studies were required to measure information on at least one outcome related to learning aligned with the recent NRC report on Education for Life and Work (Pellegrino & Hilton, 2012). This report categorized learning into three broad domains: cognitive, intrapersonal, and interpersonal. The cognitive domain includes cognitive processes and strategies, knowledge, and creativity. The intrapersonal domain includes intellectual openness, work ethic and conscientiousness, and positive core self-evaluation. The interpersonal domain includes teamwork, collaboration, and leadership.

Publication Type

To reflect the current state of digital game design, eligible studies were required to have been published between January 2000 and September 2012 in a peer-reviewed journal article. Restricting eligibility to publications in peer-reviewed journals was selected to provide consistent sampling across the diverse fields and databases covered in the literature search as outlined in the Search Strategies section below. Nonetheless, to be sensitive to any biases this may have created in our study set, we conducted extensive sensitivity analyses to assess for the possibility of publication bias, as outlined below in the Data Analysis section.

Study Site and Language

Eligible studies were those published in English (but not necessarily conducted in English or in an English-speaking country).

Effect Sizes

Eligible studies were required to report sufficient information needed to calculate both pretest and posttest effect sizes on at least one measure of learning and the variables involved in the effect sizes had to have a known direction of scoring. We use the term *pretest effect size* to indicate an effect size used to index the baseline (i.e., pretest) differences between two groups on any measure subsequently assessed at a posttest follow-up.

Search Strategies

We wanted to maximize sensitivity in our search, that is, to locate all studies that might potentially meet the eligibility criteria. Our search criteria therefore simply specified that the term *game* or *games* be included in the study abstract or title. All other search terms were deemed likely to preclude identification of potentially eligible studies. Because research on games for learning spans many fields, we searched the following hosts/databases: ISI Web of Science (SSI, SSSI), Proquest (ERIC, PsycINFO, Soc Abstracts, Social Services Abstracts), PubMed; Engineering Village (Inspec, Compendex), and IEEE Xplore. We also hand-checked the bibliographies in narrative reviews and meta-analyses.

Coding Procedures

Eligibility coding first occurred at the title level, where two research assistants independently screened all titles identified in the literature search to eliminate clearly ineligible reports (e.g., reports in non-English languages) or publications

that reported on games that were clearly irrelevant for the current study (e.g., discussion of the Olympic Games or sports injuries). Eligibility coding next occurred at the abstract level. All research assistants were first trained on a randomly selected subset of 100 abstracts, which were discussed until 100% consensus was reached with the entire group. The remaining abstracts were screened independently by two research assistants, and any disagreements were resolved by one of the authors. If there was any ambiguity about potential eligibility based on the abstract, we erred on the side of inclusivity at this stage. The final stage of eligibility coding occurred at the full-text level, in which all reports previously identified as potentially eligible at the abstract level were screened for final eligibility. At least two research assistants conducted independent full-text screening of each article, and any questions about eligibility were resolved by consensus with one of the study authors. The reason for ineligibility was recorded for each study, using the criteria outlined above.

Studies that were deemed ineligible at the full-text level were not coded further. Studies identified as eligible at the full-text level progressed to full-study coding, in which two of the study authors coded all game and nongame condition characteristics while two research assistants independently extracted information about the studies, participants, research conditions, and effects sizes. Any discrepancies in the coding were discussed in person and resolved via consensus between coders and at least one of the study authors.

Variables and Effect Size Moderators

Data were extracted for the following study characteristics and used for descriptive purposes and/or examined as potential effect size moderators.

Study Characteristics

We coded publication year, attrition between pretest and posttest measurement points, whether the study used an experimental or controlled quasi-experimental research design, location of study, whether the study had poor reporting of statistical or game-related information, and the timing at which the posttest measurement occurred.

Participant Characteristics

We coded percentage of White/non-White participants, percentage of male participants, and average age of sample.

Condition Characteristics

We measured several general characteristics related to the focal game condition in each study: duration of game, total number of game sessions, number of days elapsed between first and last game session, number of URLs provided for the game, number of screenshots provided for the game, word count of the game description, and whether the game included additional nongame instruction. In terms of game design characteristics, we measured presence of additional nongame instruction to supplement the game, sophistication of game mechanics, variety of actions in which the player engaged, social structuring of players within the game, intrinsic/extrinsic nature of the integration of learning and game

mechanics, nature of scaffolding, primary learning mechanic, visual realism, anthropomorphism, camera perspective, story relevance, and story depth. Nongame conditions were coded for comparison condition quality. Value-added comparisons were coded for the focal compared feature.

Outcome Characteristics

We coded whether the outcome was measured using an existing normed instrument, a modification of an existing instrument, or an author-developed instrument. We also coded assessments in terms of broad learning outcome domain, learning outcome discipline, and possible overalignment with the game condition.

Statistical Methods

Effect Size Metric

The outcomes of interest in the meta-analysis were measured with pretest–adjusted posttest standardized mean difference effect sizes. They were coded so that positive effect sizes represent better learning outcomes for the focal game condition of interest at the posttest follow-up time point. Pretest-adjusted posttest standardized mean difference effect sizes (d) were calculated as follow:

$$d = \left(\frac{\bar{X}_{G1} - \bar{X}_{G2}}{s_{\text{pooled}}} \right)_{\text{POST}} - \left(\frac{\bar{X}_{G1} - \bar{X}_{G2}}{s_{\text{pooled}}} \right)_{\text{PRE}},$$

where the first term is the posttest standardized mean difference effect size and the second term is the pretest standardized mean difference effect size. For each term, the numerator is the difference in means for the focal game and comparison group (using posttest means in the first term and pretest means in the second term), and the denominator is the pooled standard deviation for the scores in those groups (using the pooled posttest standard deviation in the first term and the pooled pretest standard deviation in the second term). We used this effect size metric in an attempt to provide conservative estimates of digital game effects on learning, net of pretest differences between groups on learning measures. Using a simple unadjusted posttest effect size metric would have been inappropriate given the inclusion of studies using quasi-experimental research designs where participants were not randomized to conditions.

All effect sizes were then adjusted with the small-sample correction factor to provide unbiased estimates of effect size (g) as per Hedges (1981). The small-sample corrected effect size and its standard error were calculated as follows:

$$g = \left[1 - \left(\frac{3}{4N - 9} \right) \right] * d,$$

$$SE_g = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1} * n_{G2}} + \frac{g^2}{2(n_{G1} + n_{G2})}},$$

where N is the total posttest sample size for the game and comparison groups, d is the original standardized mean difference effect size, n_{G1} is the posttest sample

size for the focal game group, and n_{G2} is the posttest sample size for the comparison group.[3] Effect size and sample size outliers were Winsorized to less extreme values (Tukey, 1977).

Some studies in the meta-analysis required cluster adjustments (Hedges, 2007) because assignment to game/comparison conditions was performed at the school or classroom level, but results were reported at the individual level and this clustering was not accounted for in the statistical analysis. Because none of the studies provided the intraclass correlations (ICCs) needed to make cluster adjustments, we made the following cluster adjustments to the standard errors of the effect sizes using a conservative estimate (based on Hedges & Hedberg, 2007) of the ICC at .20, such that

$$SE_{\text{adj}} = SE_g * \sqrt{1 + \left(M - 1\right) * \text{ICC}},$$

where SE_{adj} is the new cluster adjusted standard error, M is the number of clusters in the study, and ICC is the intraclass correlation (Higgins, Deeks, & Altman, 2008).

Several studies provided multiple effect sizes on the same learning outcome construct of interest (e.g., two different measures of mathematics learning) for the same game and comparison group combination, or, in some cases, the same study included several variants of a game condition that were all compared to a single comparison condition. This meant that effect sizes were not statistically independent. Until recently, the most statistically defensible way to handle dependent effect sizes has been to model the dependencies among effect size estimates drawn from the same study using multivariate meta-analysis techniques (Gleser & Olkin, 2009), but these methods are often difficult to implement in practice because they require information about the intercorrelations between the effect sizes, which are seldom reported in primary studies.

Therefore, we used a technique to synthesize results that allows inclusion of statistically dependent effect size estimates in a single meta-analysis and does not require information about the intercorrelation between effect sizes within studies (Hedges et al., 2010; Tanner-Smith & Tipton, 2014). With this technique, robust standard errors are used to handle the lack of statistical independence in a set of correlated effect size estimates, and no information from the source studies about outcomes need be lost to the analysis. This technique therefore permits in-depth examination of variability in the effects of digital games on learning, specifically as that variability relates to study quality, game variants, and other study characteristics.

Missing Data

To be eligible for inclusion, studies were required to provide enough information needed to estimate a pretest and posttest effect size on at least one learning outcome. Therefore, there were no missing data in the effect size outcomes of interest. Data were missing, however, on some of the coded study characteristics. Attrition data were missing for 7% of studies, race/ethnicity for 67% of studies, gender composition for 28% of studies, game duration information for 8% of

studies, and outcome measurement characteristics for 3% of studies. Because we did not have a large enough sample size to conduct any defensible imputation of missing data, we used listwise deletion for any moderator analyses that included these variables.

Analytic Strategies

Given the presumed heterogeneity in game conditions and participant samples, random effects statistical models were used for all analyses (Raudenbush, 2009). Mean effect sizes and meta-regression models using RVE were estimated using a weighted least squares approach (see Hedges et al., 2010; Tanner-Smith & Tipton, 2014, for more information). In the RVE framework, a simple model that relates the effect sizes **T** to a set of covariates in a design matrix **X** and a vector of regression coefficients **β** can be written:

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is a vector of residuals. For example, for effect size *i* in study *j*, this model can be written as follows:

$$T_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ \ldots \ + \beta_p X_{pij} + \epsilon_{ij},$$

where the effect size T_{ij} may be explained to some degree by *p* covariates $X_{1ij} \ldots X_{pij}$. The weighted least-squares estimate of $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)$ can be calculated using

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{T}),$$

where **W** is a matrix of weights. In the RVE framework, the variance of the estimate **b** of **β** can be written as

$$V^R(\mathbf{b}) = \left(\sum_{j=1}^{m} \mathbf{X}_j' \ \mathbf{W}_j \mathbf{X}_j\right)^{-1} \left(\sum_{j=1}^{m} \mathbf{X}_j' \ \mathbf{W}_j \mathbf{e}_j \mathbf{e}_j' \ \mathbf{W}_j \mathbf{X}_j\right) \left(\sum_{j=1}^{m} \mathbf{X}_j' \ \mathbf{W}_j \mathbf{X}_j\right)^{-1},$$

where for study *j*, $\mathbf{X}_j$ is the design matrix, $\mathbf{W}_j$ is the weight matrix, and $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$ is the estimated residual vector for study *j*. The random effects inverse variances weights used for the RVE analysis were calculated as

$$W_{ij} = \frac{1}{\{(V_{\cdot j} + \tau^2)[1 + (k_j - 1)]\}},$$

where $V_{\cdot j}$ is the mean of the within-study sampling variances for each study *j*, τ^2 is the estimate of the between-studies variance component, and k_j is the number of effect sizes within each study *j*.

Recent simulation studies suggest that the statistical test originally proposed by Hedges et al. (2010) has low statistical power rates unless there are large numbers of studies included in the meta-analysis (López-López, Van den

Noortgate, Tanner-Smith, Wilson, & Lipsey, 2015; López-López, Viechtbauer, Sánchez-Meca, & Marín-Martínez, 2010). Therefore, we used the *t*-test statistic suggested by López-López et al. (2010) to assess the significance of the meta-regression coefficients. All analyses were run in Stata Version 13.0. Finally, because the meta-analysis included only studies published in peer-reviewed journal articles, results could be subject to publication bias resulting from the exclusion of unpublished studies with null or negative findings (Rothstein, Sutton, & Borenstein, 2005). Therefore, we conducted sensitivity analysis to assess for the possibility of publication bias using funnel plots, Egger's regression test (Egger, Davey Smith, Schneider, & Minder, 1997) and trim and fill analysis (Duval & Tweedie, 2000).

Results

All literature searches were conducted in September 2012. Figure 1 outlines the eligibility coding for the 61,887 reports identified in the literature search. A majority of reports were initially screened out at the title level (n = 57,701). We next screened the resulting 3,141 abstracts for eligibility for coding at the full-report level. We then screened the resulting 1,040 reports in full text to determine final eligibility status. Most of the reports were ineligible for inclusion due to inadequate research designs (i.e., many were concept pieces that did not empirically examine the effect of a digital game or conduct comparisons across conditions). After screening the full-text articles, 69 unique study samples included in 70 reports from 68 journal articles ultimately met the eligibility criteria and were included in the final meta-analysis (Figure 1). These 69 study samples provided information on a total of 6,868 unique participants. Citations for the eligible journal articles are available online along with data on included studies.

Demographic and Publication Characteristics

Table 2 shows descriptive statistics for the study, participant, and outcome characteristics. As shown in Table 2, the average publication years were 2009 and 2010, with publication dates ranging from 2000 to 2012. Attrition was relatively low in most studies, with an average of only .05, which was due in large part to the immediate posttest measurement employed by many studies. Few of the studies reported the race/ethnicity of participants sufficiently to code such characteristics. For those reporting information on the gender of participant samples, roughly half of the participants were male. The average ages of participants were 12 and 13, with most participants in the seventh grade. Learning outcomes focused primarily on cognitive competencies.

Core Media Comparison and Value-Added Findings

All meta-analyses were estimated using robust variance estimates and could include multiple effect sizes from each study sample. Because the effect sizes were standardized mean difference effect sizes, confidence intervals (CIs) around mean effect sizes that include zero provide no evidence of significant differences between groups (regardless of associated effect size).

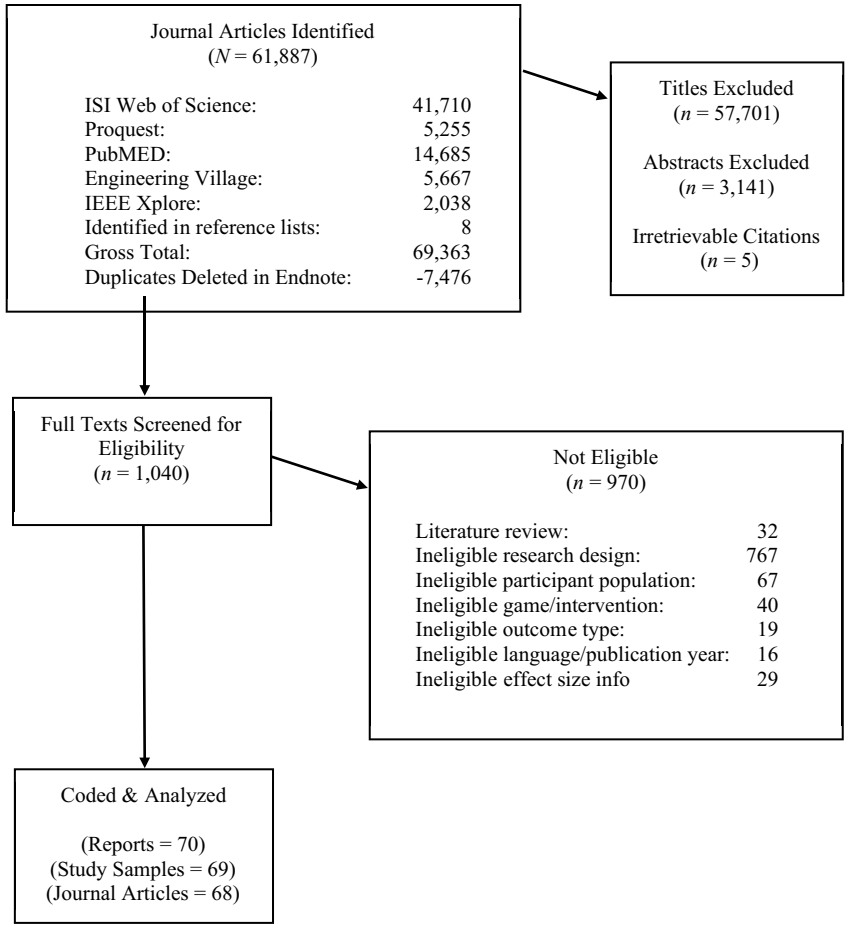

FIGURE 1. *Study identification flow diagram.*

Media Comparisons (Core Hypothesis 1)

Nongame conditions generally involved either additional "classroom as normal" time or time spent working on materials intended as representative of traditional instruction instead of playing the game. When restricting analyses to those studies comparing learning outcomes in digital game versus nongame conditions, there were 209 pairwise comparisons (i.e., effect sizes) from 57 studies across all learning domains. Among these, there were 173 effect sizes from 55 studies measuring effects on cognitive competencies, 35 effect sizes from 14 studies for intrapersonal competencies, and 1 effect size from 1 study for an interpersonal competency outcome. Table 3 shows that students in digital game conditions demonstrated significantly better outcomes overall relative to students in the nongame comparison conditions ($\bar{g}$ = 0.33, 95% CI [0.19, 0.48], τ^2 = 0.28).

TABLE 2

Descriptive statistics for study, participant, and learning outcome characteristics

	Digital game vs. non-game conditions			Digital game vs. digital game conditions			
Study characateristics	*M*	*N*	*SD*	*M*	*N*	*SD*	Range
Study context[a]							
Publication year	2009	57	2.83	2010	20	2.44	2000–2012
Attrition	0.05	53	0.10	0.05	19	0.11	0–0.40
Location of study (%)							
North America	51	57		30	20		0–100
Asia	23	57		25	20		0–100
Europe, Middle East	24	57		30	20		0–100
South America	0	57		10	20		0–100
Australia	2	57		5	20		0–100
Timing of post-test measurement (weeks)	0.66	57	3.02	0.01	20	0.03	0–21.5
Participant characteristics[a]							
% White	0.29	19	0.33	0.17	6	0.41	0–1
% Non-White	0.36	10	0.42	0.33	6	0.52	0–1
% Male	0.51	41	0.15	0.55	15	0.11	0.2–1
Average age	13.38	57	4.55	12.05	20	4.03	5–21
Learning outcome discipline[b]							
Science	13	149		17	35		0–100
Math	20	149		31	35		0–100
Literacy	18	149		0	35		0–100
Social sciences	3	149		0	35		0–100
Engineering/computer science	2	149		0	35		0–100
Psychology	31	149		31	35		0–100
General knowledge	13	149		20	35		0–100

Note. Percentages for categorical variables may not sum to 100 due to rounding error.
[a]Variables measured at study level. [b]Variables measured at outcome measure level.

The mean effect size for cognitive competencies was 0.35, indicating a significant beneficial effect of digital games on cognitive learning outcomes, relative to comparison conditions (95% CI [0.20, 0.51], $\tau^2 = 0.29$). The mean effect size for intrapersonal competencies was also 0.35, indicating a significant beneficial effect of digital games on intrapersonal learning outcomes, relative to comparison conditions ($\overline{g} = 0.35$, 95% CI [0.06, 0.65], $\tau^2 = 0.20$). Results from a

TABLE 3

Results from moderator analyses examining differences in posttest mean effect sizes for digital game versus nongame conditions.

Moderator variable	*p*	$\bar{g}$	95% Confidence interval	*n* (*k*)	τ^2
Overall learning outcomes					
All learning outcomes		0.33	[0.19, 0.48]	209 (57)	0.28
Cognitive learning outcomes		0.35	[0.20, 0.51]	173 (55)	0.29
Intrapersonal learning outcomes		0.35	[0.06,0 65]	35 (14)	0.20
Number of game sessions	*				
Single session		0.08	[−0.24, 0.39]	43 (17)	0.31
Multiple sessions		0.44	[0.29, 0.59]	166 (40)	0.22
Game includes additional nongame instruction					
Yes		0.36	[0.19, 0.52]	72 (22)	0.13
No		0.32	[0.11, 0.52]	137 (36)	0.39
Game players	*				
Single, no collaboration/ competition[a,b,c]		0.45	[0.29, 0.61]	150 (44)	0.28
Single, competitive[a,d]		−0.06	[−0.61, 0.48]	13 (4)	0.01
Single, collaborative		0.01	[−1.43, 1.44]	5 (3)	0.42
Collaborative team competition[b,d]		0.22	[−0.32, 0.76]	12 (3)	0.00
Multiplayer/MMO[c]		−0.05	[−0.31, 0.21]	29 (7)	0.16
Game type					
Adding points/badges		0.53	[0.27, 0.79]	64 (17)	0.28
More than points/badges		0.25	[0.08, 0.42]	145 (40)	0.27
Variety of game actions					
Small		0.35	[0.08, 0.61]	67 (19)	0.28
Medium		0.43	[0.25, 0.62]	109 (27)	0.22
Large		0.40	[0.25, 0.55]	176 (46)	0.24
Intrinsic/extrinsic type	*				
Not fully intrinsic		0.33	[−0.09, 0.74]	19 (10)	0.38
Intrinsic		0.19	[−0.02, 0.41]	110 (24)	0.25
Simplistically intrinsic		0.49	[0.27, 0.71]	80 (23)	0.27
Scaffolding	*				
Success/fail/points[e]		0.26	[0.05, 0.47]	118 (25)	0.29
Answer display		0.40	[−0.07, 0.88]	11 (3)	0.00
Enhanced scaffolding		0.48	[0.18, 0.78]	36 (15)	0.30
Teacher-provided scaffolding[e]		0.58	[0.20, 0.96]	8 (4)	0.03

(continued)

TABLE 3 (CONTINUED)

Moderator variable	*p*	$\bar{g}$	95% Confidence interval	*n* (*k*)	τ^2
Visual realism	*				
Schematic[f]		0.48	[0.13, 0.82]	52 (12)	0.34
Cartoon		0.32	[0.13, 0.50]	80 (20)	0.17
Realistic[f]		−0.01	[−0.34, 0.32]	36 (13)	0.33
Anthropomorphism	*				
Low/none[g]		0.37	[0.19, 0.56]	125 (39)	0.32
Medium[g]		0.04	[−0.26, 0.33]	48 (9)	0.21
High		0.55	[−0.58, 1.69]	6 (3)	0.49
Camera view	*				
First person (FPS, POV)		0.12	[−0.33, 0.57]	15 (8)	0.31
Over the shoulder/overhead tracking[h]		−0.02	[−0.35, 0.31]	32 (10)	0.29
Third person[h]		0.48	[0.30, 0.66]	140 (32)	0.26
Story relevance	*				
None		0.44	[0.16, 0.71]	46 (17)	0.28
Irrelevant[i]		0.63	[0.33, 0.94]	75 (11)	0.27
Relevant[i]		0.17	[−0.03, 0.37]	88 (29)	0.26
Story depth					
None[j]		0.44	[0.16, 0.71]	46 (17)	0.28
Thin[k]		0.47	[0.27, 0.67]	98 (22)	0.23
Medium[j,k]		−0.03	[−0.31, 0.24]	44 (13)	0.15
Thick		0.36	[−0.43, 1.15]	21 (5)	0.59

Note. MMO = massively multiplayer online; FPS = first-person shooter; POV = point of view. $\bar{g}$ = mean posttest effect sizes adjusted for pretest differences between groups. *n* = number of effect sizes. *k* = number of unique study samples. τ^2 = between studies variance component. The 95% confidence intervals were estimated using robust variance estimates. Because the effect sizes were standardized mean difference effect sizes, confidence intervals for mean effect sizes that include zero provide no evidence of differences between the game and nongame groups. Asterisks are used to indicate significant differences in mean effect sizes by game characteristic, per coefficients from meta-regression models with robust variance estimates. Superscripts denote pairwise differences between indicated game characteristics.
*$p < .05$.

meta-regression model using robust variance estimates provided no evidence of a difference in the magnitude of the mean effect sizes for the cognitive and intrapersonal learning outcomes ($b = 0.03$, 95% CI [−0.32, 0.38]). Only one study provided an effect size on an interpersonal learning outcome (not shown separately in Table 3), which was positive and favored the digital game conditions but was not significantly different from zero ($\bar{g} = 0.25$, 95% CI [−0.28, 0.77]). Overall, the results therefore indicated that digital games improved students' learning outcomes by approximately 0.3 standard deviations relative to typical instruction

TABLE 4

Posttest mean effect sizes for enhanced design variants of digital games versus equivalent standard versions of those digital games

Design variant	$\bar{g}$	95% Confidence interval	*n* (*k*)	τ^2
All enhanced designs versus all standard versions	0.34	[0.17, 0.51]	40 (20)	0.10
Enhanced scaffolding designs versus equivalent standard versions	0.41	[0.18, 0.64]	20 (9)	0.11
Collaborative social designs versus equivalent standard versions	0.24	[−0.33, 0.81]	6 (3)	0.03
Competitive social designs versus equivalent standard versions	0.33	[−1.13, 1.78]	6 (3)	0.41
Providing/situating context versus equivalent standard versions	0.32	[−0.53, 1.16]	3 (3)	0.11
Interface enhancement designs versus equivalent standard versions	0.39	[−0.13, 0.90]	3 (2)	0.00
Extended game play design versus equivalent standard versions	0.70	[−0.84, 2.25]	1 (1)	—
Enhanced scaffolding + competition (2 × 2 combination design)	−0.22	[−1.01, 0.56]	1 (1)	—

Note. $\bar{g}$ = mean posttest effect sizes adjusted for pretest differences between groups, *n* = number of effect sizes, *k* = number of unique study samples; 95% confidence intervals estimated using robust variance estimates. Because the effect sizes were standardized mean difference effect sizes, confidence intervals for mean effect sizes that include zero provide no evidence of differences between the game and nongame groups.

conditions, and these effects were similar in magnitude across the cognitive and intrapersonal competencies domains.

Value-Added Comparisons (Core Hypothesis 2)

Value-added comparisons measure the efficacy of a standard version of a game relative to an enhanced version of that game augmented to test a theoretical design proposition (Mayer, 2011). For the purposes of comparison, conditions were identified as including the standard version of a game or an enhanced version of that game. Table 4 shows the results from this analysis. Overall, the comparison of all enhanced versions versus standard versions showed a significant positive effect size for the enhanced designs ($\bar{g}$ = 0.34, 95% CI [0.17, 0.51]). This comparison involved 40 effect sizes estimated from 20 studies and included all categories of enhanced designs.

As part of this hypothesis, we had also planned to explore specific categories of value-added comparisons in terms of the focal compared feature represented in the enhancement. The only category with substantial representation turned out to be enhanced scaffolding, which included 20 effect sizes drawn from 9 studies.

Enhanced scaffolding was defined broadly to include personalized scaffolding, intelligent agents, adapting game experiences to student needs or interests, and revised game structuring targeted at emphasizing the learning mechanic. Specific comparisons of enhanced scaffolding demonstrated a significant overall effect size of similar magnitude to the overall value-added findings ($\bar{g}$ = 0.41, 95% CI [0.18, 0.64]). None of the other categories independently demonstrated mean effect sizes that were significantly different from zero, but this is likely influenced by the small number of effect sizes in each comparison.

Moderator Analyses of General Study Characteristics

Play Duration (Moderator Hypothesis 1a)

As shown in Table 3, game conditions involving multiple game play sessions demonstrated significantly better learning outcomes than nongame control conditions, but there was no evidence that game conditions involving single game play sessions were different from nongame control conditions. Furthermore, effects were significantly smaller when games were played in one game session versus more than one session (b = −0.37, p = .03, 95% CI [−0.70, −0.04]). In terms of overall game play duration, on average, media comparison game interventions involved 11 sessions over the course of 38 days for a total duration of 347 minutes. Results from a meta-regression model provided no evidence of an association between total game play duration and effect size magnitude (b = 0.00, p = .99, 95% CI [−0.0005, 0.001]).

Because the effect of the absolute duration of an intervention might differ widely depending on game characteristics, we reestimated the meta-regression models after controlling for visual realism, anthropomorphism, variety of game actions, viewpoint, story relevance, and story depth. The purpose was to examine whether the differences in observed effects across categories would remain after controlling for those other game characteristics. The relationship between single-session versus multiple-session comparisons remained statistically significant, and the relationship between total duration and effect size magnitude remained nonsignificant.

Additional Instruction (Moderator Hypothesis 1b)

Many studies included game conditions with additional nongame instruction (e.g., students participating in relevant classroom work in addition to game play). As shown in Table 3, there was no evidence that effects were different depending on whether or not the game conditions included additional nongame instruction (b = 0.04, 95% CI [−0.23, 0.31]).

Player Configuration (Moderator Hypothesis 1c)

In terms of player grouping structure, effects of digital games on learning outcomes were largest for those game conditions using single noncollaborative/noncompetitive play. In fact, this was the only group that demonstrated significant learning gains, although this could be due to the small number of conditions and effect sizes in each of the other categories (see Table 3). Moreover, average effects were significantly larger for games with single players (with no formal

collaboration or competition) relative to those using single/competitive play (b = 0.69, p < .001, 95% CI [0.39, 0.99]), collaborative team competitions (b = 0.29, p = .03, 95% CI [0.02, 0.56]), or multiplayer/MMOs (b = 0.49, p = .001, 95% CI [0.21, 0.78]). Games with collaborative team competition, however, produced significantly larger effects than those using single/competitive players (b = 0.40, p = .001, 95% CI [0.17, 0.63]).

Because player grouping structure might be correlated with other game characteristics, we reestimated the meta-regression models while controlling for visual realism, anthropomorphism, variety of game actions, viewpoint, story relevance, and story depth. Results from that model indicated that after controlling for those other game characteristics, games with single noncollaborative/noncompetitive players still exhibited significantly larger mean effect sizes than those with single competitive players (b = 0.71, p = .01, 95% CI [0.23, 1.18]) but no longer had significantly different effect sizes than games with collaborative team competitions or multiplayer games. Collaborative team competitions still exhibited significantly larger effect sizes than those with single/competitive players (b = −0.48, p = .02, 95% CI [−0.86, −0.10]).

Moderator Analyses of Game Mechanics Characteristics

Sophistication of Mechanics (Moderator Hypothesis 2a)

The first category of broad sophistication of game design focuses on relatively rudimentary games involving the mere addition of points and/or badges to schoollike tasks. As shown in Table 3, these games were associated with a 0.53 standard deviation improvement in learning outcomes. Games in the second category could include those rudimentary aspects, but they also included mechanics, scaffolding, and/or situating context beyond those rudimentary aspects. This second category of games was associated with a 0.25 standard deviation improvement in learning. Although these results suggest that the average effect was largest for rudimentary games, results from a meta-regression model including a dummy indicator for the game type indicated no significant differences in the mean effect size across the two categories (b = 0.28, p = .07, 95% CI [−0.02, 0.57]).

Variety of Player Actions (Moderator Hypothesis 2b)

The next section of Table 3 presents results in terms of the variety of game actions in which the player engaged during the game (i.e., small, medium, or large). Small variety includes simple games, such as *Tetris*, where players engage in a relatively small variety of actions on screens that change little over the course of the game. Note that graphics sophistication is not the key variable here; a virtual reality space where players simply rotate objects or explore the structure of a protein would also be considered small in terms of variety. Medium variety includes game conditions with a modest variety of actions with multiple manners of interacting with and exploring an environment (e.g., *Zoombinis*), or game conditions with multiple simple games that together provide a moderate variety of game actions. Large variety includes game environments, such as *Quest Atlantis* or *SimCity*, with a large variety of game actions and means of interacting with the environment.

Overall, effects on learning were strongest for game conditions with medium or large varieties of game actions. Effects were somewhat smaller for those games with a small variety of actions, but there was no evidence that the mean effect sizes across these three categories were different from each other. Results were similar after controlling for visual realism, anthropomorphism, viewpoint, story relevance, and story depth.

Intrinsic Integration (Moderator Hypothesis 2c)

We also classified game conditions based on the integration of the primary learning mechanic and the primary game play mechanic (cf. Habgood & Ainsworth, 2011; Kafai, 1996). The learning mechanics can be defined as the mechanics and interactions intended to support players in learning the target learning outcomes. The game mechanics can be defined as the mechanics and interactions ostensibly designed for engagement and progress in the game.

Interestingly, there was only one game with a fully extrinsic relationship between core learning and game mechanics. Instead, many of the games were *simplistically intrinsic* or *intrinsic*. The simplistically intrinsic category included simple game designs with only a single mechanic that served as both the learning mechanic and the game mechanic (e.g., *Tetris* or a game where answering questions or problems is the only mechanic). By comparison, the intrinsic game conditions involved fully intrinsic designs where the primary learning mechanic is integrated into the core atoms of the game mechanic in a more complex structure. Games labeled as *not fully intrinsic* included game conditions involving a mix of extrinsically and intrinsically integrated learning mechanics (as well as the single fully extrinsic study).

Although the mean effect size was slightly larger for games using simplistically intrinsic designs relative to those using intrinsic or not fully intrinsic designs, there was no evidence that the mean effect sizes were significantly different across these three categories (Table 3). Results were similar even after controlling for the visual realism, anthropomorphism, variety of game actions, viewpoint, story relevance, and story depth game characteristic variables.

Scaffolding (Moderator Hypothesis 2d)

We compared four categories of scaffolding. Game conditions in the lowest category provide scaffolding only in terms of indicating success/failure or number of points earned by the player. The next category includes scaffolding that additionally displays the answer/solution in some manner after an error. The next category provides enhanced scaffolding beyond simply indicating success/failure and displaying the correct answer (e.g., intelligent agents or adapting scaffolding to past performance). The highest category (in terms of adaptiveness of the scaffolding) involves scaffolding provided by the teacher. As shown in Table 3, although results were relatively similar across the scaffolding categories, the effect on learning outcomes was significantly larger for games where the teacher provided scaffolding relative to those games using simple success/failure/points ($b = 0.33$, $p = .05$, 95% CI [0.00, 0.66]). This result could be considered a comparison of the lowest level of scaffolding (points/success) to the highest levels of scaffolding (teachers) in terms of adaptiveness and specificity. This difference in

effects across categories remained statistically significant after controlling for the visual realism, anthropomorphism, variety of game actions, viewpoint, story relevance, and story depth of the game.

Moderator Analyses of Visual and Narrative Game Characteristics

Visual Realism (Moderator Hypothesis 3a)

Visual realism focuses on the graphical realism of the game environment. The schematic category includes schematic, symbolic, and text-based games with overall simplistic graphical elements. The cartoon category includes games with nonrealistic shading or forms (e.g., nonrealistic forms of characters or objects), often in a two-dimensional format. The realistic category includes games with realistic shading/forms or real pictures, often in a three-dimensional format. Effects were significantly larger for schematic than realistic games ($b = 0.45$, $p = .03$, 95% CI [0.05, 0.84]), but this difference was attenuated to marginal significance ($b = 0.44$, $p = .09$, 95% CI [−0.95, 0.06]) after controlling for anthropomorphism, variety of game actions, viewpoint, story relevance, and story depth game characteristic variables.

Anthropomorphism (Moderator Hypothesis 3b)

We coded anthropomorphism as the degree to which the player, nonplayable characters, and environmental entities in the game have human features or perform humanlike movements. For an entity to be considered relevant for the purposes of coding, attention to the entity must be important for successful gameplay. The low/none category includes either few or no anthropomorphic entities or features. The medium category includes approximately equal numbers of anthropomorphic and nonanthropomorphic entities. The high category includes a majority of anthropomorphic entities and features and anthropomorphic qualities closer to human. As shown in Table 3, effects were significantly larger for games using low/no anthropomorphizing compared to those using medium levels of anthropomorphizing ($b = 0.33$, $p = .047$, 95% CI [0.004, 0.67]), but this difference was attenuated to nonsignificance ($b = 0.22$, $p = .24$, 95% CI [−0.15, 0.60]) after controlling for visual realism, variety of game actions, viewpoint, story relevance, and story depth game characteristic variables.

Perspective (Moderator Hypothesis 3c)

Camera perspective is the camera viewpoint through which players interact with the game. If the game included cut-scenes in which the player did not control actions, these cut-scenes were not considered when coding for camera perspective. The third person viewpoint presents noncamera-based views (e.g., *Tetris*) or third-person camera views in which the camera is neither presented as being through the eyes of the player nor presented from the perspective of the player (e.g., *Tycoon City: New York*). The over the shoulder or overhead tracking viewpoint presents the world in three dimensions through a moving camera that follows the player's avatar but is not presented through the eyes of the player (e.g., *Super Mario 3D*). The first person viewpoint presents the game world as if through the avatar's eyes (e.g., *Portal*). As shown in Table 3, effects were significantly larger for games using third person viewpoints relative to those using over the

shoulder/overhead tracking viewpoints ($b = 0.37$, $p = .02$, 95% CI [0.05, 0.67]), but this difference was attenuated to nonsignificance ($b = 0.25$, $p = .36$, 95% CI [−0.80, 0.30]) after controlling for visual realism, anthropomorphism, variety of game actions, story relevance, and story depth game characteristic variables.

Story Relevance (Moderator Hypothesis 3d)

Story relevance assesses whether or not the narrative is relevant to the learning mechanic. Story relevance is different from the relationship between the game mechanic and learning mechanic (intrinsic vs. extrinsic) because it deals specifically with the story rather than the game mechanic. A story about analyzing scientific data in a game that requires applying math skills to graphs of experimental data would be relevant (e.g., *McClarin's Adventures*), for example, but a story about killing zombies in a game that requires solving simple math problems (e.g., *Zombie Division*) would not be relevant. We coded in terms of no story, an irrelevant story, or a relevant story. As shown in Table 3, effects were significantly larger for game conditions using irrelevant story lines compared to those using relevant story lines ($b = 0.46$, $p = .01$, 95% CI [0.12, 0.81]), but this difference was attenuated to nonsignificance ($b = 0.15$, $p = .56$, 95% CI [−0.37, 0.67]) after controlling for visual realism, anthropomorphism, variety of game actions, viewpoint, and story depth game characteristic variables.

Story Depth (Moderator Hypothesis 3e)

Story depth categorizes the extent of the story. Thin depth involves only setting, scenery, or context. Medium depth involves some evolving story over the course of the game. Thick depth includes a rich evolving story over the course of the game. Results showed that games with no story or thin story depth both had significantly larger effects relative to those with medium story depth ($b = 0.47$, $p = .02$, 95% CI [0.10, 0.84], and $b = 0.51$, $p = .003$, 95% CI [0.18, 0.84], respectively). These differences were unchanged when controlling for visual realism, anthropomorphism, variety of game actions, viewpoint, and story relevance. Games with thick story depth did not demonstrate significant effects on learning relative to nongame conditions or other levels of story depth, but only five studies included thick story depth, limiting the power of comparisons involving games with thick story depth.

Contextualization (Moderator Hypothesis 3f)

One issue with individual analyses of visual realism, anthropomorphism, camera viewpoint, story relevance, and story depth is that these characteristics are likely intercorrelated. For example, we found significant correlations at the $p < .001$ level between viewpoint and visual realism ($r = .60$), visual realism and anthropomorphism ($r = .54$), visual realism and story relevance ($r = .47$), and visual realism and story depth ($r = .49$). Given these confounds, we constructed an aggregate measure of contextualization from the above components. The contextualization score was calculated as the sum of five game context items: view point (1 = *third person*, 2 = *over the shoulder/overhead tracking*, 3 = *first person*), visual realism (1 = *schematic*, 2 = *cartoon*, 3 = *realistic*), anthropomorphism (1 = *low*, 2 = *medium*, 3 = *high*), story relevance (1 = *none*, 2 = *irrelevant*,

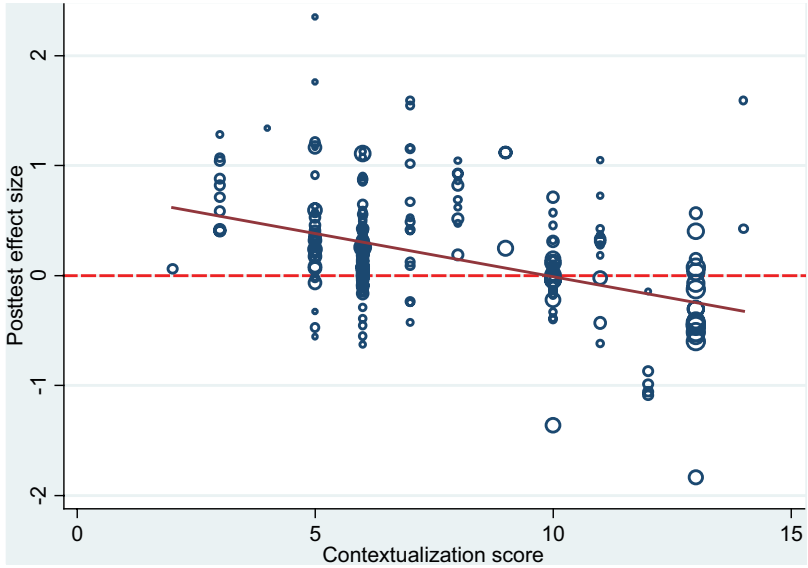

FIGURE 2. *Scatter plot of pretest-adjusted posttest effect sizes and overall contextualization aggregate score for digital game versus nongame conditions (media comparisons).*
Note. Each effect size shown proportionate to its weight in the meta-analysis. Slope coefficient from meta-regression with robust variance estimation $b = -0.07$ ($p = .01$, 95% CI [−0.12, −0.01]).

3 = *relevant*), and story depth (1 = *thin*, 2 = *medium*, 3 = *thick*). Missing values (and other codes such as *unknown* or *mixed*) received a value of 0 in the summative contextualization score. Figure 2 shows the relationship between this summative contextualization score and the effect of digital games on learning outcomes. Results from a meta-regression model examining the magnitude of this relationship showed a significant negative relationship ($b = -0.07$, $p = .01$, 95% CI [−0.12, −0.01]), indicating that increased contextualization was correlated with smaller effects on learning outcomes.

Research Characteristics in Value-Added and Media Comparisons

Comparison Condition Quality (Moderator Hypothesis 4a)

Comparison condition quality tracks the equivalence of the control condition to the game condition in terms of the focal comparison (i.e., the manipulation the authors indicated as the primary focus). We coded comparison conditions as (a) sham or irrelevant activities, (b) weak comparisons, (c) medium comparisons representing rough equivalents of typical classroom approaches but not representing tightly controlled matches, (d) strong comparisons designed and optimized as clearly viable alternatives but still not tightly controlled matches, and (e) excellent comparisons representing direct analogs controlling for all but the focal variables. Results indicated that comparison condition quality had a significant relationship

to effect sizes in the media comparison analyses ($b = -0.15$, $p = .01$, 95% CI [−0.27, −0.04]). As shown in Table 5, if we restrict the meta-analysis only to those studies with a medium or better comparison condition, that would reduce the mean effect size from 0.33 to 0.28. Comparison condition quality was not associated with effect size magnitude in the value-added analyses.

Condition Reporting (Moderator Hypothesis 4b)

Condition reporting was coded in terms of word count and number of screenshots for game conditions. Many studies provided minimal information about the game conditions. Table 5 shows the results from analyses restricted to studies based on the word count of the game description. Overall, there were minimal differences in effects when we filtered based on word counts of the game descriptions for media comparison or value-added analyses. Number of screenshots, however, was significantly correlated with effect sizes for the game versus nongame conditions in the media comparison analyses ($b = 0.05$, $p = .02$, 95% CI [0.01, 0.09]). Indeed, if the meta-analysis was restricted to only those studies that included one or more screenshots, the mean effect size increased from 0.33 to 0.37 for media comparison analyses. Filtering based on number of screenshots did not significantly change effect size for value-added analyses.

Methods Reporting (Moderator Hypothesis 4c)

We coded each study subjectively in terms of insufficiency of reporting of methods and analyses. Specifically, we coded whether studies reported clearly inappropriate statistical analyses (e.g., analyzing cluster-randomized trial data at the individual unit of analysis with no adjustment for clustering) and/or demonstrated serious omissions in the reporting of methods or statistical analyses (e.g., omission of standard deviations or sample sizes, confusion between posttest and pretest–posttest change scores). Although sufficiency of reporting methods and analyses was not significantly associated with effect size magnitude in the meta-regression models for the media comparison or value-added analyses, it is noteworthy that the mean effect size was reduced substantially if we restricted the media comparison analyses to only those studies with unflawed reporting of their methods or analyses.

Assessment Overalignment (Moderator Hypothesis 4d)

We coded each study in terms of subjective overalignment of assessment with the game tasks. Specifically, we applied a binary code to indicate whether studies used learning outcome measures that were partially or entirely overaligned with the learning tasks included in the game conditions themselves (e.g., an English proficiency test of vocabulary questions that included the same vocabulary questions appearing in the digital quiz game under investigation). This characteristic was not significantly associated with effect size in the meta-regression models for the media comparison or value-added analyses, and there were minimal differences in effects when we filtered based on overalignment with outcome.

Assessment Type (Moderator Hypothesis 4e)

Assessments were categorized as preexisting normed instruments, modifications of preexisting instruments, or author-developed instruments. Results

TABLE 5

Posttest mean effect sizes for digital game versus nongame conditions for all learning outcomes by study quality variables

Study quality variable	*Significance*	$\bar{g}$	95% Confidence interval	*n* (*k*)	τ^2
All media comparisons (MC)		0.33	[0.19, 0.48]	209 (57)	0.28
MC comparison condition quality	*				
Medium or better		0.28	[0.12, 0.43]	175 (48)	0.29
MC sufficient condition reporting: screenshots	*				
Includes 1 or more screenshots		0.37	[0.16, 0.57]	79 (33)	0.34
MC sufficient condition reporting: word count					
Includes 500 or more words description of conditions		0.36	[0.15, 0.58]	76 (28)	0.35
MC reporting of methods and analyses					
No flaws in reporting methods or analyses		0.15	[−0.15, 0.46]	44 (15)	0.33
MC overalignment of assessed outcome with task					
No apparent overalignment		0.30	[0.16, 0.43]	181 (48)	0.21
MC assessment type					
Author developed instrument only		0.33	[0.11, 0.56]	97 (24)	0.32
Modification of an existing instrument only		0.48	[0.17, 0.80]	19 (10)	0.12
Preexisting normed instrument only		0.40	[0.22, 0.58]	89 (37)	0.22
MC research design					
Quasi-experimental design only		0.43	[0.22, 0.63]	96 (25)	0.22
Experimental design only		0.17	[0.004, 0.33]	113 (32)	0.20
All value-added comparisons		0.34	[0.17, 0.51]	40 (20)	0.10
Comparison condition quality					
Medium or better		0.34	[0.17, 0.51]	40 (20)	0.10
Sufficient condition reporting: no. of screenshots					
Includes 1 or more screenshots		0.33	[0.11, 0.54]	25 (15)	0.12
Sufficient condition reporting: word count					
Includes 500 or more words description of conditions		0.32	[0.15, 0.49]	25 (13)	0.07
Reporting of methods and analyses					
No flaws in reporting methods or analyses		0.20	[−0.00, 0.39]	22 (11)	0.09

(continued)

TABLE 5 (CONTINUED)

Study quality variable	*Significance*	$\bar{g}$	95% Confidence interval	*n* (*k*)	τ^2
Overalignment of assessed outcome with task					
No apparent overalignment of assessed outcome		0.25	[0.06, 0.43]	30 (17)	0.06
Assessment type					
Author-developed instrument only		0.27	[−0.02, 0.56]	16 (10)	0.09
Modification of an existing instrument only		0.46	[−0.60, 1.52]	3 (3)	0.16
Preexisting normed instrument only		0.33	[0.09, 0.56]	20 (11)	0.12
Research design					
Quasi-experimental design only		0.50	[0.02, 0.99]	8 (6)	0.13
Experimental design only		0.28	[0.10, 0.47]	32 (14)	0.09

Note. Asterisks used to indicate significant differences in mean effect sizes by quality characteristic, per coefficients from meta-regression models with robust variance estimates.
*$p < .05$.

indicated no significant differences in effect size magnitude across assessment types for media comparison or value-added analyses, although the mean effect sizes varied slightly and nonsignificantly from 0.33 for author-developed instruments to 0.40 for preexisting normed instruments for media comparison analyses.

Study Design (Moderator Hypothesis 4f)

Research design was not associated with effect size magnitude for value-added analyses. Research design was not significantly associated with effect size magnitude for media comparison analyses but was marginal ($b = -0.26$, $p = .051$, 95% CI [−0.51, 0.001]). The average effect size was notably larger in the quasi-experimental studies than the randomized controlled trials (see Table 5). Given this, we explored the bivariate correlations between study design and other game characteristics for the media comparison analyses. The correlations below were significant at the $p < .05$ level. In terms of study characteristics, randomized trials were more likely to use collaborative team competitions ($r = .23$). In terms of game design, randomized trials were less likely to include game mechanics that simply added points and badges ($r = -.47$), more likely to include a greater variety of game actions ($r = .29$), less likely to include intrinsic games ($r = -.22$), less likely to use answer display ($r = -.21$), and less likely to involve teaching provided scaffolding ($r = -.22$). In terms of visual and narrative contextualization, randomized trials were more likely to have thick story lines ($r = .23$) and less likely to use third-person viewpoints ($r = -.21$). Thus, the trend of smaller observed effects among the randomized controlled trials may be due to variations in the types of games used.

Restricting Analyses With Multiple Study Quality Characteristics

We consider comparison condition quality, sufficient condition reporting, sufficient reporting of methods and analyses, and overalignment of assessment to be quality variables to which all studies should be held accountable (i.e., study design–independent). We consider assessment type and research design to be study design–dependent quality variables in the sense that they must be weighed against other research choices (which we clarify in the Discussion section).

For media comparison analyses, only four studies met all study design–independent filters, and only two studies met all study quality filters. Synthesizing results for those comparisons yields nonsignificant mean effect sizes ($\bar{g}$ = 0.02, 95% CI [−0.72, 0.76], and $\bar{g}$ = −0.11, 95% CI [−3.25, 3.04], respectively). For value-added analyses, only six studies met all of the study-design-independent filters and only two studies met all filters. Synthesizing results for those comparisons yields nonsignificant mean effect sizes ($\bar{g}$ = 0.21, 95% CI [0.02, 0.40], and $\bar{g}$ = 0.11, 95% CI [−0.71, 0.92], respectively). These results must be interpreted with extreme caution given the small number of studies and effect sizes available for the analysis (and the uncertainty in these estimates is reflected in the wide confidence intervals).

Publication Bias

Finally, related to research quality, we also explored the possibility of publication bias within our sample. Figure 3 shows the funnel plots with pseudo 95% confidence limits for the media comparison (top) and value-added analyses (bottom). There were no obvious asymmetries in either funnel plot by outcome type, and results from Egger regression tests provided no evidence of small study effects/bias for media comparison (b = 0.53, p = .36) or value-added (b = −0.90, p = .33) analyses. Furthermore, results from trim and fill analyses yielded no trimmed or filled data points for either plot, providing additional support that it is unlikely that these findings suffered from publication/small-study bias.

Discussion

Overall, results indicated that digital games were associated with a 0.33 standard deviation improvement relative to nongame comparison conditions. Thus, digital games conditions were on average more effective than the nongame instructional conditions included in those comparisons. These results generally confirm the overall findings from prior meta-analyses on the effects of games on learning (Sitzmann, 2011; Vogel et al., 2006; Wouters et al., 2013). Findings from the present meta-analysis do diverge slightly from the finding in Wouters et al. (2013) that game conditions and nongame instructional conditions did not differ in terms of motivation outcomes. In the current study, the intrapersonal learning outcome domain not only included motivation but also included intellectual openness, work ethic and conscientiousness, and positive core self-evaluation. Thus our findings do not necessarily conflict with those of Wouters et al. (2013) but rather suggest that game conditions support overall improvements in intrapersonal learning outcomes relative to nongame instructional conditions.

In terms of value-added comparisons, augmented game designs were associated with a 0.34 standard deviation improvement in learning relative to standard

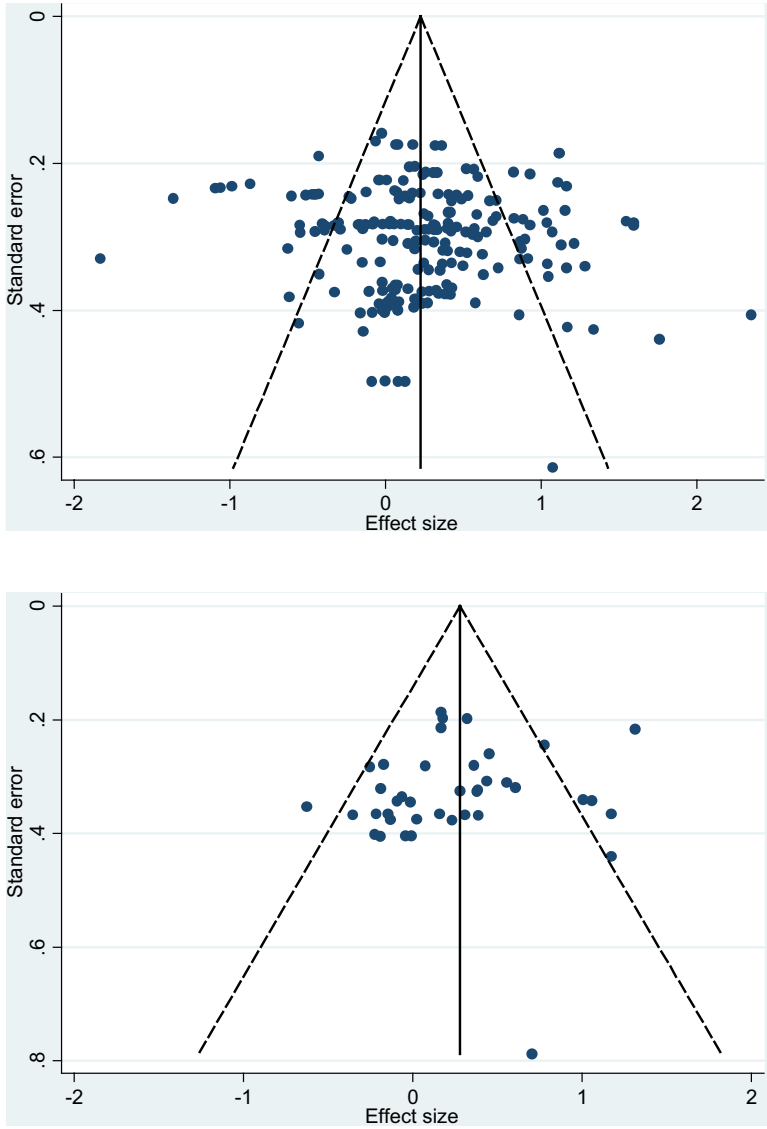

FIGURE 3. *Funnel plots with pseudo 95% confidence limits for media comparisons (top) and value-added comparisons (bottom).*

versions. This, along with the largely overlapping confidence intervals around the mean effect sizes in the media comparison and value-added analyses, suggests that the effects for the media comparison and value-added comparisons were similar in magnitude. This result highlights that the design of an intervention is

associated with as large an effect as the medium of an intervention. Although this conclusion may appear to be common sense, the role of design is often de-emphasized in debates over whether digital games are better or worse than traditional instruction. The value-added findings empirically demonstrate the importance of the role of design beyond medium. Although too few value-added comparisons met eligibility requirements to support moderator analyses of design features, the findings underscore the need to carefully consider moderator analyses of differences in design across game conditions in the media comparisons, in terms of better understanding the role of design as well as in terms of interpreting the nature and import of what is being compared.

Moderator Analyses of General Study Characteristics

Play Duration (Moderator Hypothesis 1a)

Similar to results from prior meta-analyses, we found that (a) game conditions involving multiple game play sessions demonstrated significantly better learning outcomes than nongame control conditions and (b) game conditions involving single game play sessions did not demonstrate different learning outcomes than nongame control conditions. In our analysis that focused on total game play duration (i.e., as a continuous moderator variable), however, we found no evidence of a relationship between total duration and effects on learning outcomes.

It is worth noting that the constituent studies involved largely equivalent amounts of treatment time between experimental and comparison conditions (rather than simply comparing treatment time increases in experimental conditions while holding control conditions constant). The findings may therefore reflect a benefit of spaced learning as compared to massed learning in game contexts (cf. Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; McDaniel et al., 2013). Longer play durations may enhance learning but only when sessions are adequately spaced. Alternatively, it is possible that the impact of total duration was masked by game conditions that were longer than needed to achieve the observed improvement on the assessments. Games were played for an average of 347 minutes (or almost 6 hours). Deeper assessments of student learning should thus be investigated in future research.

Additional Instruction (Moderator Hypothesis 1b)

Additional nongame instruction was not associated with larger or smaller effects for game conditions in media comparisons. These findings diverge from Sitzmann (2011) and Wouters et al. (2013), who found that supplemental nongame instruction supported learning. Both Sitzmann (2011) and Wouters et al. (2013) may have used a more stringent definition for "additional instruction." In the present meta-analysis, game conditions were coded as including additional nongame instruction (whether integrated or not) if players were exposed to a learning context that was likely to provide them with additional topic-relevant information (e.g., spending days in typical classroom instruction). Based on Wouters et al.'s (2013) examples, it is possible that only studies that explicitly stated that players received additional domain-relevant instruction were coded as such. This might suggest that additional teaching or activities specifically designed to supplement game content as part of an integrated experience can increase

learning but unintegrated supplemental instruction is unlikely to contribute to larger gains. Another important clarification is that Sitzmann (2011), Wouters et al. (2013), and the current meta-analysis do not include interaction of players with informal sites or communities on the Internet as "nongame instruction" (e.g., World of Warcraft community forums or Wiki Game support sites). Research has demonstrated the importance and power of the argumentation and learning that occur on these sites (e.g., Steinkuehler & Duncan, 2008). Thus, findings from the three meta-analyses do not analyze that important context of learning and participation around games.

Player Configuration (Moderator Hypothesis 1c)

When controlling for game characteristics, single-player games without competition and collaborative team competition games outperformed those from single-player games with competition. These findings partly parallel those of Wouters et al. (2013), who found that collaborative play was generally more effective than individual play. Our findings, however, suggest that collaborative games may not be generally more effective for learning than single-player games but that competitive single-player structures may be least effective. This explanation would align with research on motivation and learning (e.g., Bandura, 1997; Pintrich, 2003; Schunk, 1991). The motivational support of self-efficacy for certain students in a single-player competitive structure is necessarily a failure to support other students (because one student's gain necessitates another student's loss).

This comparison highlights, however, the challenges of aggregating interventions across studies and games in terms of the potential to mask important instructional variables. Although meta-analysis as a method assumes commensurability across studies, confounding variables are inevitably present when synthesizing aggregate findings from multiple studies. In the current analysis, it is possible that the goals for individual versus group game conditions differed in a manner that contributed to the observed overarching pattern. Indeed, single-player games might have induced players to pursue a goal of attaining the highest possible score, for example, whereas collaborative games might have induced players to adopt or test maximizing strategies for team member roles. Interpretation thus requires careful consideration of possible underlying variables and mechanisms of change. We will return to this challenge in the Caveats and Limitations section.

Moderator Analyses of Game Mechanics Characteristics

The comparison of broad design sophistication in media comparisons (Moderator Hypothesis 2a) demonstrated that simple gamification as well as more sophisticated game mechanics can prove effective. To clarify this finding, future research and analyses should explore whether or not the simple gamification studies (e.g., games that simply add contingent points and badges to learning activities) more frequently focus on lower order learning outcomes as compared to studies with more sophisticated game mechanics. Regardless, these results support the proposal that simple gamification can prove effective for improving certain types of learning outcomes (cf. Lee & Hammer, 2011; Sheldon, 2011). These findings parallel those observed for the variety of game actions (Moderator

Hypothesis 2b), showing equivalent learning outcomes across all levels of action variety in media comparison studies.

The present meta-analysis is largely silent with regard to intrinsic versus extrinsic design (Moderator Hypothesis 2c) because only one study involved a fully extrinsic condition. Regarding the nature of scaffolding (Moderator Hypothesis 2d), each category of scaffolding demonstrated significant effects on learning relative to nongame control conditions, but higher levels of scaffolding were associated with higher relative learning outcomes than lower levels of scaffolding. Enhanced scaffolding also showed significant effects on learning outcomes in the value-added analyses. These findings provide a productive foundation for ongoing work on enhancing scaffolding in games (e.g., Barzilai & Blau, 2014).

Moderator Analyses of Visual and Narrative Game Characteristics

Several visual and narrative game characteristics (Moderator Hypotheses 3a–3e) were intercorrelated. An aggregate contextualization variable created from these game characteristics (Moderator Hypothesis 3f) demonstrated a small but significant negative relationship with learning gains overall in media comparisons. This result parallels the findings of Wouters et al. (2013), which showed that schematic games were more effective than cartoon-like or realistic serious games and supports the trend those authors observed that games with no narrative might be more effective than games with narratives.

On the surface, these findings contradict research and theory highlighting the value of situating learning in context (e.g., Bransford et al., 2000). One possible interpretation is that rich narratives and visual complexity distract students from the intended learning content or provide alternative goals within the game that do not support improvement on the assessed outcome measures. This interpretation would speak to the need for game designers and education researchers to collaborate on designs to keep game graphics, environments, and narratives optimally aligned with assessed learning objectives.

A second possible interpretation focuses on the nature of the assessments in the constituent studies. Almost all the studies analyzed in this report involved immediate posttests focusing on lower order learning outcomes. The arguments for situating learning in context focus on developing a deep, durable, integrated understanding that students can apply across contexts (essentially the opposite of an immediate focused posttest). This interpretation highlights the importance of including assessments designed to measure deeper understanding in future research. Such a shift in assessment would align with theoretical proposals indicating that the greatest strengths of digital games as a medium involve their affordances for supporting higher order cognitive, intrapersonal, and interpersonal learning objectives (e.g., Gee, 2007; Squire, 2011).

The visual and narrative features of games are also envisioned as potentially creating a "time for telling" about lower level concepts in a meaningful and compelling context. A third possible interpretation of our findings from this perspective is that our own coding rules may not have captured the critical relationships between narratives and learning in terms of time for telling about lower order learning objectives. Specifically, we coded the relevance of narratives in terms of

relevance to the learning mechanic rather than assessment content. Thus, relevant narratives may have helped contextualize the learning mechanic in the game play but failed to create a time for telling about lower level concepts in a meaningful manner in terms of the assessed learning objectives.

A fourth possible interpretation also focuses on our coding system. We coded narrative in terms of relevance and thickness, but perhaps the critical features of narratives are whether they are engaging, high-quality, or accessible, regardless of thickness or relevance. Some thin narratives are incredibly engaging, whereas some thick narratives may be dull. Additionally, poorly designed thick narratives might be difficult for students to understand. Similar questions could be framed in terms of the value of visual sophistication versus visual clarity or visual engagement. The amount of information reported about the game contexts was minimal in many of the constituent studies, restricting the ways in which we were able to code visual and narrative characteristics, but clearly much room remains for exploring the relationships between contextualization and learning.

Research Characteristics in Value-Added and Media Comparisons

Few studies met all four study design–independent quality variables for the research quality moderator analyses (Moderator Hypotheses 4a–4d) in value-added or media comparisons, supporting claims that methodological rigor needs to be improved in research on games for learning. That said, results from moderator analyses indicated that few study quality variables (design-independent or design-dependent) influenced the effects of digital games on learning outcomes in the media comparison or value-added analyses (Moderator Hypotheses 4a–4f). This provides additional confidence in our overall effect estimates and suggests that findings were not unduly biased by individual study quality variables. Further discussion (provided below), is merited, however, for one design-independent variable (control condition quality) and both design-dependent variables (assessment type and research design).

Control Condition Quality

Restricting the meta-analysis to only those studies with medium or better comparison condition quality (thus weeding out "straw man" comparisons) reduced the effect size from 0.33 to 0.28 (but remained significant). These findings further underscore the importance of design (and careful reporting of that design) for both game and nongame conditions (cf. Young et al., 2012). Media comparison research often highlights medium while placing less emphasis on the design of the game and control conditions. Many of the media comparison studies in the present report, for example, provided only sparse descriptions of game or control interventions. As research on games begins to focus more on design, researchers will need to provide thicker descriptions of conditions to support informed comparisons across studies.

Assessment Type

There are trade-offs between research questions of interest and the availability of preexisting normed instruments. Although preexisting assessments can clearly enhance confidence in research quality, these instruments exist only for certain

outcomes. Furthermore, the present meta-analysis found no evidence of a relationship between assessment type (i.e., preexisting normed instrument, modification of a preexisting instrument, or author-developed instrument) and effect size magnitude. The present meta-analysis also found no evidence of a relationship between effect sizes and potential overalignment of assessments. Given the aforementioned trade-offs and our null result concerning the impact of normed instruments on effect sizes, we propose that requiring research to rely exclusively on preexisting normed instruments would unnecessarily limit digital games research. This issue is particularly relevant for the outcome types most desirable from the perspective of 21st-century skills and preparedness (for which normed assessments are scarce). Researchers should thus be encouraged to choose appropriate assessments based on learning goals but should report reliability and validity information for author-created or -modified instruments.

Research Design

Although there were no significant differences in average effects across randomized and controlled quasi-experimental designs in the present meta-analysis, the observed effects were notably smaller in the studies using randomized designs. Post hoc correlational analyses showed, however, that differences in game characteristics between games in studies using randomized versus quasi-experimental designs might partially account for effect size differences across study designs. Furthermore, randomized designs preclude many research questions and populations. We therefore argue that researchers should carefully weigh the benefits of experimental designs in light of fundamental issues of ecological validity, authenticity, and specific requirements of the research questions under exploration. In studies where quasi-experimental designs are implemented, researchers must provide more substantial information about the group attributes and account for those attributes in analyses.

Caveats and Limitations

This section raises three issues for consideration. The first involves commensurability, which should be considered when interpreting this (or any) meta-analysis. Meta-analyses assume that the included pairwise comparisons represent relatively standardized or homogenous conditions. In practice, this is not the case even in settings that might appear highly homogeneous, such as medical research. Jüni, Witshci, Bloch, and Egger (1999), for example, described these hazards in great detail in their article in the *Journal of the American Medical Association*. Commensurability poses even greater challenges when aggregating studies of learning and education, where variations across contexts, interventions, and approaches are more extreme.

Thus, although meta-analyses aggregate findings within categories that sound highly generalizable, the included research conditions do not fill or equally represent the entire domain suggested by the categories. Neither this nor any meta-analysis accounts for all possible design approaches or qualities of implementation. Future research should not be limited, therefore, to the highest performing game characteristics identified in the current meta-analysis. Alternative designs for low-performing game characteristics should be investigated if those characteristics are

considered critical to learning goals. We argue that this implication is particularly salient regarding our findings for visual and narrative contextualization, where overarching research highlights the importance of situating learning in context to support deeper understanding, but the findings of this meta-analysis underscore potential design and alignment challenges.

In addition to commensurability of game conditions, there are commensurability issues for the nongame comparison conditions, which generally represented typical instructional approaches rather than optimized learning activities in the constituent studies. The findings of the media comparison analyses should thus not be interpreted as suggesting that game-based instruction is superior to all learning experiences that could be designed within traditional media; rather, the findings suggest that the game-based experiences analyzed in these studies were superior to the traditional nongame approaches implemented in the constituent studies. We therefore urge against simplistic quotations of findings suggesting that games universally outperform nongame learning approaches. The results and comparisons are more complex and must be acknowledged as such.

The second issue concerns inclusion, which is related to commensurability. Meta-analyses include distinct cross sections of studies (as is true for any type of review; cf. Young et al., 2012). As shown in Table 1, Vogel et al.'s (2006) and Sitzmann's (2011) meta-analyses included simulations, for example, and less than 50% of the studies from Wouters et al. (2013) were eligible in the present meta-analysis (with publication date and research designs accounting for most differences). Furthermore, although many important studies focusing on design have been conducted in the learning sciences, games research, and other fields, not all these studies met the eligibility criteria for inclusion in this particular meta-analysis (often based on the requirement of experimental or quasi-experimental designs involving pretest–posttest measurements, sufficient reporting for calculation of effect sizes, or eligible comparison conditions). This is important to note because research conducted from some epistemological paradigms, particularly sociocultural paradigms, can be relatively incompatible with current assessment practices and experimental designs. The current meta-analysis therefore includes only a cross section of research on games, and eligibility should not be conflated with contribution or value. We need to leverage the findings across studies, regardless of their eligibility for inclusion in the current analyses, as we move forward in exploring the role of design to leverage the affordances of games for learning.

The third issue concerns assessments. Higher order cognitive, intrapersonal, and interpersonal processes and skills prove more challenging to measure accurately and reliably than do lower order cognitive skills and rote knowledge. As a result, research on games has generally focused on lower order cognitive skills, rote knowledge, and immediate posttests. The NRC report on education for life and work in the 21st century, however, emphasizes a more distributed focus across outcomes, if not a complete reversal in emphasis. Furthermore, proponents of digital games for learning (e.g., Gee, 2007; Squire, 2011) propose that the greatest strengths of digital games as a medium involve their affordances for supporting higher order cognitive, intrapersonal, and interpersonal learning objectives. Assessments that yield reliable and valid scores of higher order processes and

skills would also facilitate further research at sociocultural and situated grain sizes of the overarching activity structure and community, as well as over much longer longitudinal time frames of months or years rather than hours or days, which are the grain sizes and time frames underlying the greatest strengths of games for learning (cf. Young et al., 2012). For all these reasons, ongoing development and research should focus more heavily on accurate and reliable assessment of higher order learning outcomes.

Role of Design and Final Thoughts

To date, much experimental and quasi-experimental research on games and learning has focused on media comparisons. The present meta-analysis suggests that games as a medium can indeed support productive learning. Furthermore, the results of the present meta-analysis parallel the conclusions of the NRC report on laboratory and inquiry activities (Singer, Hilton, & Schweingruber, 2005) in highlighting the key role of design beyond medium. Thus, harkening back to the Clark/Kozma debates of the 1980s and 1990s about the relative importance of studying medium versus design (e.g., Clark, 1994; Kozma, 1994), games as a medium definitely provide new and powerful affordances, but it is the design within the medium to leverage those affordances that will determine the efficacy of a learning environment. We now need to leverage findings on games from across methodological paradigms, regardless of their eligibility for inclusion in the current analyses, to conduct situated empirical analyses that consider design in terms of interactions among player goals, game affordances, pedagogy, teaching objectives, and curricular content. Our findings expand on and reinforce Young et al.'s (2012) findings that we should "stop seeking simple answers to the wrong question" (p. 84). We should thus shift emphasis from proof-of-concept studies ("Can games support learning?") and media comparison analyses ("Are games better or worse than other media for learning?") to cognitive-consequences and value-added studies exploring how theoretically driven design decisions influence situated learning outcomes for the broad diversity of learners within and beyond our classrooms.

Notes

This work was supported by the Games Learning and Assessment Lab-Research (GlassLab-Research) grant from The Bill and Melinda Gates Foundation to SRI International. This article serves as the final full report for that work. Special acknowledgments to the team that assisted in coding and screening, including Shara Bellamy, Jamie Eldredge, Lauren Kissenger, Kaitlin Reynolds, Kasia Steinka-Fry, Marriah Vinson, Eric Wilkey, and Stephanie Zuckerman.

[1]In addition to these three quantitative meta-analyses, two other recent endeavors are noteworthy. First, Girard, Ecalle, and Magnan (2013), analyzed 11 digital games with the intent of conducting a meta-analysis but did not end up calculating and reporting the meta-analysis results because they felt that not enough studies had been conducted that looked specifically at "serious games" or "true digital games." Second, Ke (2009) conducted a qualitative review that provides interesting insights and syntheses of research on game-based learning without statistically summarizing effect sizes.

[2]When analyzing learning outcomes associated with moderator variables, learning outcomes for focal experimental conditions are calculated relative to the comparison condition

outcomes in the associated studies. In interest of parsimony, this clause is omitted from the statement of the moderator hypotheses.

[3]Given the way in which results were reported in the primary studies included in the meta-analysis, we were unable to estimate standard errors using the formula for effect sizes derived from repeated-measures analysis of variance (see Borenstein, 2009, p. 227, Equation 12.21). Indeed, very few studies included information on the correlation between the pretest and posttest measures, a quantity that is needed to estimate the standard error of effect sizes from repeated-measures analysis of variance models. Among the eight studies that reported the pretest–posttest correlations, they ranged widely (0.15–0.88), and thus we deemed it inappropriate to impute a common value across the remaining studies. Nevertheless, we conducted sensitivity analyses using standard error estimates using this range of plausible values for the pretest–posttest correlation, and results were substantively unchanged from those reported herein.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S., & Warren, S. (2009). Transformational play as a curricular scaffold: Using videogames to support science education. *Journal of Science Education and Technology*, *18*, 305–320. doi:10.1007/s10956-009-9171-5

Barab, S. A., Zuiker, S., Warren, S., Hickey, D., Ingram-Goble, A., Kwon, E.-J., . . . Herring, S. C. (2007). Situationally embodied curriculum: Relating formalisms and contexts. *Science Education*, *91*, 750–782. doi:10.1002/sce.20217

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education*, *70*, 65–79. doi:10.1016/j.compedu.2013.08.003

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*, 2nd edition (pp. 221-235). New York, NY: Russell Sage Foundation.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn*. Washington, DC: National Academies Press.

Bransford, J. D., Sherwood, R. D., Hasselbring, T. S., Kinzer, C. K., & Williams, S. M. (1990). Anchored instruction: Why we need it and how technology can help. In R. J. Spiro (Ed.), *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 115–141). New York, NY: Routledge.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. doi:10.3102/0013189X018001032

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi:10.1037/0033-2909.132.3.354

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research & Development*, *42*, 21–29. doi:10.1007/BF02299088

Dickey, M. D. (2006). Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. *Educational Technology Research & Development*, *54*, 245–263. doi:10.1007/s11423-006-8806-y

Dieterle, E. (2009). Neomillennial learning styles and River City. *Children, Youth & Environments*, *19*, 245–278. Retrieved from http://www.jstor.org/action/showPublication?journalCode=chilyoutenvi

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98. doi:10.1080/01621459.2000.10473905

Echeverria, A., Barrios, E., Nussbaum, M., Amestica, M., & Leclerc, S. (2012). The atomic intrinsic integration approach: A structured methodology for the design of games for the conceptual understanding of physics. *Computers & Education*, *59*, 806–816. doi:10.1016/j.compedu.2012.03.025

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. doi:10.1136/bmj.315.7109.629

Federation of American Scientists. (2006). *Summit on educational games: Harnessing the power of video games for learning*. Washington, DC: Author. Retrieved from http://informalscience.org/images/research/Summit_on_Educational_Games.pdf

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*, 850–855. doi:10.1111/j.1467-9280.2007.01990.x

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). New York, NY: Palgrave Macmillan.

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*, 207–219. doi: 10.1111/j.1365-2729.2012.00489.x

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York, NY: Russell Sage Foundation.

Green, C. S., & Bavelier, D. (2006). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology*, *32*, 1465-1478. doi:10.1037/0096-1523.32.6.1465

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18*, 88–94. doi:10.1111/j.1467-9280.2007.01853.x

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.3102/10769986006002107

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. doi:10.3102/1076998606298043

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, *22*(10), 1–15.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. doi:10.1002/jrsm.5

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259.

Higgins, J. P. T., Deeks, J. J., & Altman, D. G. (2008). Special topics in statistics. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic review of interventions* (pp. 481-529). Hoboken, NJ: Wiley.

Hines, P. J., Jasny, B. R., & Mervis, J. (2009). Adding a T to the three R's. *Science*, *323*, 53. doi:10.1126/science.323.5910.53a

Honey, M. A., & Hilton, M. (Eds.). (2010). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.

Jüni, P., Witshci, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, *282*, 1054–1060. doi:10.1001/jama.282.11.1054

Kafai, Y. B. (1996). Learning design by making games: Children's development of strategies in the creation of a complex computational artifact. In Y. B. Kafai & M. Resnick (Eds.), *Constructionism in practice: Designing, thinking and learning in a digital world* (pp. 71–96). Mahwah, NJ: Erlbaum.

Kafai, Y. B. (2006). Playing and making games for learning: Instructionist and constructionist perspectives for game studies. *Games and Culture*, *1*, 36–40. doi:10.1177/1555412005281767

Kafai, Y. B., Quintero, M., & Feldon, D. (2010). Investigating the "'Why'" in Whypox: Casual and systematic explorations of a virtual epidemic. *Games and Culture*, *5*, 116–135. doi:10.1177/1555412009351265

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. Ferdig (Ed.), *Handbook of Research on Effective Electronic Gaming in Education* (pp. 1–32). New York: IGI Global.

Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City: A multi-user virtual environment. *Journal of Science Education and Technology*, *16*, 99–111. doi:10.1007/s10956-006-9038-y

Killingsworth, S. S., Levin, D. T., & Saylor, M. M. (2011). Analyzing action for agents with varying cognitive capacities. *Social Cognition*, *29*, 56–73.

Klopfer, E., Scheintaub, H., Huang, W., Wendel, D., & Roque, R. (2009). The simulation cycle: Combining games, simulations, engineering and science using StarLogo TNG. *E-Learning and Digital Media*, *6*, 71–96. doi:10.2304/elea.2009.6.1.71

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. *Educational Technology Research & Development*, *42*(2), 7–19. doi:10.1007/BF02299087

Lee, J. J., & Hammer, J. (2011). Gamification in education: What, how, why bother? *Academic Exchange Quarterly*, *15*(2), 1–5.

Lim, C. (2008). Global citizenship education, school curriculum and games: Learning Mathematics, English and Science as a global citizen. *Computers & Education*, *51*, 1073–1093. doi:10.1016/j.compedu.2007.10.005

López-López, J. A., Van den Noortgate, W., Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2015). *Assessing meta-regression models for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation*. Manuscript in preparation.

López-López, J. A., Viechtbauer, W., Sánchez-Meca, J., & Marín-Martínez, F. (2010, July). *Comparing the performance of alternative statistical tests for moderators in mixed-effects meta-regression models*. Paper presented at the 5th annual meeting of the Society for Research Synthesis Methodology, Cartagena, Spain.

Mahajan, N., & Woodward, A. L. (2009). Seven-month-old infants selectively reproduce the goals of animate but not inanimate agents. *Infancy*, *14*, 667–679. doi:10.1080/15250000903265184

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and*

instruction. Cognitive and affective process analyses (Vol. *3*, pp. 223–253). Hillsdale, NJ: Lawrence Erlbaum.

Martinez-Garza, M., Clark, D. B., & Nelson, B. (2013). Digital games and the U.S. National Research Council's science proficiency goals. *Studies in Science Education*, *49*, 170–208. doi:10.1080/03057267.2013.839372

Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 281–305). Charlotte, NC: Information Age.

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology*, *39*, 1417–1432. doi:10.1037/a0032184

Pellegrino, J., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington DC: National Academies Press.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*, 667–686. doi:10.1037/0022-0663.95.4.667

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: Russell Sage Foundation.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: John Wiley.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231. doi:10.1080/00461520.1991.9653133

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences*, *4*, 321–354. doi:10.1207/s15327809jls0403_3

Sheldon, L. (2011). *The multiplayer classroom: Designing coursework as a game*. Boston, MA: Cengage.

Singer, S., Hilton, M. L., & Schweingruber, H. A. (2005). *America's lab report: Investigations in high school science*. Washington, DC: National Academies Press.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, *64*, 489–528. doi:10.1111/j.1744-6570.2011.01190.x

Squire, K. (2011). *Video games and learning: Teaching and participatory culture in the digital age*. New York, NY: Teachers College Press.

Squire, K. D., & Jan, M. (2007). Mad city mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology*, *16*, 5–29. doi:10.1007/s10956-006-9037-z

Squire, K. D., & Klopfer, E. (2007). Augmented reality simulations on handheld computers. *Journal of the Learning Sciences*, *16*, 371–413. doi:10.1080/10508400701413435

Steinkuehler, D., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology*, *17*, 530–543. doi:10.1007/s10956-008-9120-8

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations and a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*, 13–30. doi:10.1002/jrsm.1091

Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2013). The comparative effectiveness of outpatient treatment for adolescent substance abuse: A meta-analysis. *Journal of Substance Abuse Treatment*, *44*, 145–158. doi:10.1016/j.jsat.2012.05.006

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, *4*, 169–187. doi:10.1002/jrsm.1070

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van der Meij, H., Albers, E., & Leemkuil, H. (2011). Learning from games: Does collaboration help? *British Journal of Educational Technology*, *42*, 655–664. doi:10.1111/j.1467-8535.2010.01067.x

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*, 229–243.

Wilson, S. J., Tanner-Smith, E. E., Lipsey, M. W., Steinka-Fry, K., & Morrison, J. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school-aged children and youth. *Campbell Systematic Reviews*, *8*. doi:10.4073/csr.2011.8

Wouters, P., Paas, F., & van Merriënboer, J. J. M. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, *78*, 645–675. doi:10.3102/0034654308320320

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*, 249–265. doi:10.1037/a0031311

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. doi:10.3102/0034654312436980

Authors

DOUGLAS B. CLARK is a professor of the learning sciences and science education at Vanderbilt University, Box 230, 230 Appleton Place, Nashville, TN 37203-5721; e-mail: *doug.clark@vanderbilt.edu*. He earned his doctorate at UC Berkeley. He researches students' conceptual change processes and approaches for scaffolding those processes in games and other technology-rich environments. He is currently principal investigator (PI) of the National Science Foundation (NSF) Discovery Research K-12 (DRK12) *Enhancing Games With Assessment and Metacognitive Emphases* and the Department of Education IES *Explanation and Prediction Increasing Gains and Metacognition* grants that focus on scaffolding students' conceptual change processes in digital game–based environments as well as developing approaches for analyzing game play data in real time for formative evaluation and adaption of student's experiences in those environments. He is also CoPI on the NSF cyberlearning grant *Fostering Computational Thinking in Middle Schools Through Scientific Modeling and Simulation*. He was PI of the exploratory NSF DRK12 grant *Scaffolding Understanding by Redesigning Games for Education* and was on the leadership team for the NSF Centers for Learning and Teaching grant *Technology Enhanced Learning in Science*.

EMILY E. TANNER-SMITH is a research assistant professor at the Peabody Research Institute and Department of Human and Organizational Development at Vanderbilt

University, Box 0181, 230 Appleton Place, Nashville, TN 37203-5721; e-mail: *e.tanner-smith@vanderbilt.edu*. She earned her doctorate at Vanderbilt University. She is a research methodologist with emphasis in systematic reviewing and meta-analysis, and her substantive areas of expertise include the social epidemiology, prevention, and treatment of adolescent delinquency and substance use. Her recent research appears in *Campbell Systematic Reviews, Journal of Substance Abuse Treatment, Journal of Youth and Adolescence, Prevention Science*, and *Research Synthesis Methods*.

STEPHEN S. KILLINGSWORTH is a postdoctoral scholar in the Department of Teaching and Learning at Vanderbilt University, Box 230, 230 Appleton Place, Nashville, TN 37203-5721; e-mail: *s.killingsworth@vanderbilt.edu*. He earned his doctorate in cognitive psychology at Vanderbilt. He is currently conducting research under the Enhancing Games With Assessment and Metacognitive Emphases and the Explanation and Prediction Increasing Gains and Metacognition grants. His research focuses on experimental and individual differences approaches to investigating visual cognition and memory in game-based learning in order to improve both game and assessment design.