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

Douglas B. Clark, Emily E. Tanner-Smith, Stephen Killingsworth   March 2014

Purpose

In 2006, the Federation of American Scientists issued a widely publicized report stating that games as a medium offer a powerful new educational tool (FAS, 2006). 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), as have reports by the National Research Council (Honey & Hilton, 2010; NRC, 2009). These reports acknowledged, however, the sparseness of systematic evidence for the efficacy of games as learning tools. The present meta-analysis synthesized research on digital games to systematically examine their efficacy for learning. We focused on research published between 2000 and 2012 in light of the dramatic evolution of digital games for learning over the past decade.

Meta-Analysis Approach

A meta-analysis is the systematic synthesis of quantitative results from a collection of studies (Borenstein, et al., 2009) focused on a given topic. Part of the systematic approach in a meta-analysis is to document the decisions that are made regarding the collection of the articles and the steps of the analysis. In a meta-analysis, articles are included based on pre-defined criteria and not due to favorable results found in the article or familiarity with certain authors. This can help to remove some of the bias and subjectivity that would result from a less systematic review.

Meta-analysis quantifies results by using effect sizes. Effect sizes are a measure of the difference between two groups. In the case of an intervention, an effect size can be thought of as a measure of the (standardized) difference between the control group and the treatment group, thereby providing a measure of the effect of the intervention. Effect sizes are not the same as statistically significant differences that are typically reported and determined through the use of inferential statistics, such as t-tests or analysis of variance (ANOVAs). A research study, for example, could have a statistically significant finding, but the effect of that difference could be minimal. Thus the effect size allows researchers to determine the magnitude of the impact of an intervention, not just whether or not the intervention made a difference. For example, an effect size of 1.00 would be interpreted as a difference of one standard deviation between the two groups being compared. Another way of interpreting a one standard deviation effect size would be moving a student at the 50th percentile before the intervention to the 84th percentile after the intervention.

SRI Education

Developed by SRI International with funding from the Bill & Melinda Gates Foundation.

The current meta-analysis employs a recently developed statistical technique for robust variance estimation in meta-regression (Hedges, Tipton, & Johnson, 2010). 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., 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 the full report for more details and Tanner-Smith & Tipton, 2013, for a discussion).

Search and Inclusion Criteria

Digital game. Eligible studies were required to include at least one comparison of a digital game versus a non-game condition and/or at least one comparison of an augmented game design versus an equivalent standard game design (but these two types of comparisons were always analyzed separately).

Participants. Eligible participant samples included students in grades K-16, ages 6 to 25.

Research designs. Only studies using randomized controlled trial and quasi-experimental research designs were eligible for inclusion.

Learning outcomes. Eligible studies were required to measure information on at least one eligible outcome related to "learning" aligned with the recent NRC report on Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century (Pellegrino & Hilton, 2012).

Publication type. To reflect the current state of digital game design, eligible studies were required to have been published between 2000 and 2012 in a peer-reviewed journal article.

Study site and language. Eligible studies were those published in English.

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.

Literature Search

All literature searches were conducted in September 2012. We wanted to maximize sensitivity in our search. Our database search criteria therefore simply specified that the term game or games needed to be included in the abstract or title. All other potential search terms were deemed likely to eliminate otherwise eligible studies. Because research on games for learning spans many fields including engineering, computer science, medical, natural sciences, and social sciences, 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. The database search identified 61,887 potential reports. Many reports were 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 in the meta-analysis 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.

Overarching Media-Comparison and Value-Added Findings

Hypothesis 1: Students in digital game conditions will outperform students in non-game conditions in terms of learning outcomes. Fifty-seven studies included comparisons of digital game interventions versus other non-game instructional conditions (i.e., media comparisons). Overall, results indicated that digital games were associated with a .33 standard deviation improvement relative to control conditions, even after adjusting for baseline differences in achievement between groups. Thus, the analyses show that digital games conditions were on average more effective than the non-game instructional conditions included in those comparisons.

Hypothesis 2. Games with theoretically augmented designs for learning will outperform standard versions of those games. Twenty studies included comparisons of augmented versions versus standard versions of digital games (i.e., value-added comparisons). Overall, results indicated that augmented game designs were associated with a .37 standard deviation improvement in learning relative to standard versions, even after adjusting for baseline differences in achievement between groups. This finding highlights the importance of design in learning outcomes. Furthermore, the largely overlapping confidence intervals around the mean effect sizes from the media-comparison and value-added analyses suggest that the findings for the media-comparison and value-added analyses were similar in magnitude. This suggests that the design of an intervention is associated with as large an effect as the medium of an intervention. Although this finding may appear common-sense, the role of design is often de-emphasized in debates over whether digital games are "better" or "worse" than traditional instruction. It is critical to consider this finding when interpreting the media-comparison analyses.

General Study Characteristics in Media Comparisons

Hypothesis 3a: Comparisons involving more than a single game-play session, as well as comparisons involving longer play durations, will be associated with better learning outcomes relative to non-game conditions. As found by Wouters et al. (2013), we found that (a) game conditions involving multiple game-play sessions demonstrated significantly better learning outcomes than non-game control conditions and (b) game conditions involving single game-play sessions did not demonstrate significantly different learning outcomes than non-game control conditions. We then conducted analyses of game-play duration (treating total gameplay duration as a continuous moderator variable), but we found no evidence of a consistent correlation between total duration and effects on learning outcomes.

Taken together, these findings may reflect a memory benefit of spaced learning as compared to massed learning in game contexts. Longer play durations may thus enhance learning, but only when sessions are adequately spaced. Alternatively, the distinction in findings may simply suggest that (a) a single session is not sufficient but (b) the multiple-session game conditions were being played longer than needed to achieve relevant learning outcomes in most studies. Games were played for an average of 347 minutes (or almost 6 hours). It may be that students were able to learn relatively quickly the concepts relevant for the assessments used in those studies. From this perspective, deeper assessments of student learning would be highly desirable for future research.

Hypothesis 3b: Game conditions with non-game instruction will outperform game conditions without non-game instruction relative to non-game conditions. Additional non-game instruction was not significantly 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 non-game instruction supported learning. One possible explanation for this difference involves how "additional instruction" was coded across meta-analyses, suggesting that additional

teaching or activities specifically designed to supplement game content as part of an integrated experience can increase learning, but that unintegrated supplemental teaching on a topic is unlikely to contribute to larger gains.

Hypothesis 3c: Collaborative game conditions will outperform single-player game conditions relative to non-game conditions. When controlling for game characteristics, gains from single-player games without competition and gains from collaborative team competition games exceeded those from single-player games with competition. These findings partly parallel the findings of Wouters et al. (2013), but may elaborate upon their findings. Wouters et al. found that collaborative play was generally more effective than individual play. Our findings suggest that collaborative games may not be more effective for learning than single player games, but instead suggest that games with competitive single-player structures are least effective.

Game Mechanics and Contextualization Characteristics in Media Comparisons

We predicted that *more sophisticated game mechanics, increased variety of player actions, intrinsic integration of the game mechanic and learning mechanic, and more specific/detailed scaffolding will be related to larger effects on learning outcomes relative to non-game conditions (Hypotheses 4a-4d)*. The comparison of broad design sophistication (Hypothesis 4a) demonstrated that simple gamification as well as more sophisticated game mechanics can prove effective. 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. These findings parallel those regarding variety of game actions (Hypothesis 4b), in which significant learning outcomes relative to non-game control conditions were observed for all levels of action variety with no significant differences between them.

The present meta-analysis is largely silent with regard to intrinsic versus extrinsic design (Hypothesis 4c) due to the fact that only one study involved a fully extrinsic condition. Regarding the nature of scaffolding (Hypothesis 4d), each category of scaffolding demonstrated significant effects on learning relative to non-game 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.

Based on the findings of prior meta-analyses in this field, we predicted that *visual realism, anthropomorphism, camera perspective, story relevance, and story depth will be related to smaller effects on learning outcomes relative to non-game conditions (Hypothesis 5a-5e) as will overall aggregate contextualization (Hypothesis 5f)*. As mentioned, we also investigated and controlled for the relationships between these features. In addition, given the trends observed in prior research, we predicted that greater overall contextualization will be related to smaller effects on learning outcomes relative to non-game conditions (Hypotheses 5f). The influence of individual visual and narrative game characteristics (Hypotheses 5a-5e) proved highly intercorrelated. An aggregate contextualization variable created from these game features (Hypothesis 5f) demonstrated a small but significant negative relationship with learning gains overall. Our findings parallel the findings of Wouters et al. (2013), (1) showing that schematic games were more effective than cartoon-like or realistic serious games and (2) supporting the non-significant trend observed in their study suggesting that games with no narrative might be more effective than games with narratives.

Research Quality Characteristics in Value-Added and Media Comparisons

To explore research quality characteristics more deeply, we predicted that *comparison condition quality, sufficient condition reporting, sufficient reporting of methods and analyses, over-alignment of assessment with game, assessment type, and study design will be related to learning outcomes in value-added and media comparisons (Hypotheses 6a-6f)*. Few studies met all four study-design-independent quality variables for Hypotheses 6a-6d, supporting claims that overall rigor needs to be increased in research on games for learning. That said, results from moderator analyses indicated that few study quality variables (design-independent or design-dependent) were highly correlated individually with the effects of digital games on learning outcomes in the media-comparison or value-added analyses (Hypotheses 6a-6f). This provides additional confidence in our effect estimates and suggests that findings were not unduly biased by individual study quality variables.

Caveats and Limitations

Meta-analyses assume that the included pairwise comparisons (effect sizes) represent relatively standardized or homogenous conditions. In actuality, this is not the case even in settings that might appear highly standardized. Even meta-analyses of medical research, for example, involve difficult balancing acts in terms of commensurability. Jüni, Witshci, Bloch, and Egger (1999) described these hazards in great detail in their *Journal of the American Medical Association* article. Commensurability issues pose even greater challenges when aggregating studies of learning and education, where variations across contexts, interventions, and approaches are more extreme.

The present meta-analysis explored only the specific game designs and methodological choices included in the constituent studies. Thus, while meta-analyses aggregate research conditions into 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 other meta-analysis can thus account for all possible design approaches or the implementation quality of those approaches.

We therefore do not suggest that future research and design should focus only on the characteristics and mechanics that outperformed others in this meta-analysis. Instead, if designs around a specific characteristic demonstrated lower learning outcomes, then other designs should be investigated if that characteristic is considered critical. We argue that this implication is particularly salient regarding our findings for visual and narrative contextualization, where overarching research supports situating cognition in terms of transfer and deeper understanding (c.f., Bransford, Brown, & Cocking, 2000), but the findings of this meta-analysis underscore challenges in terms of design implementation.

In addition to commensurability of game conditions, there are also commensurability issues for non-game comparison conditions. The studies in the present meta-analysis generally compared targeted game interventions to traditional or typical instructional approaches rather than to optimized learning activities. Thus, the findings of the media-comparison analyses should 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 non-game approaches *implemented* in the included studies. All forms of media have particular affordances and constraints that must be considered in the design of high-quality instruction. Digital games and traditional instruction necessarily vary on many dimensions. We therefore urge against simplistic quotations of findings suggesting that games universally outperform non-game learning approaches. The results and comparisons are more complex and should be acknowledged as such.

Role of Design and Final Thoughts

Much of the research to date on digital games has focused on proof-of-concept studies and media comparisons. The present meta-analysis highlights the importance of questions that ask not *if* but *how* games can support learning. More specifically, the results of the present meta-analysis parallel those of the recent NRC report on laboratory and inquiry activities (Singer, Hilton, & Schweingruber, 2005). Design, rather than medium alone, predicts learning outcomes. Research on games and game-based learning 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 value-added comparisons and cognitive-consequences studies exploring how theoretically-driven design decisions influence learning outcomes for the broad diversity of learners within and beyond our classrooms.

References

Borenstein, M., Hedges, L., Higgins, J. & Rothstein, H. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Bransford, J. D., A. L. Brown, and Cocking, R.R., eds. (2000). *How People Learn.* Washington, D.C., National Academy Press.

Federation of American Scientists. (2006). *Report: Summit on Educational Games: Harnessing the Power of Video Games for Learning.* Washington, D.C.

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.

Hines, P. J., Jasny, B. R., & Merris, J. (2009). Adding a T to the three R's. *Science, 323,* 53.

Honey, M. A., & Hilton, M. (Eds.). (2010). *Learning Science through Computer Games and Simulations.* National Research Council. Washington, DC: National Academy 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*(11), 1054–1060.

National Research Council. (2009). *National Research Council Workshop on Games and Simulations.* October 6-7, 2009, Washington, D.C.

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.

Tanner-Smith, E. E., & Tipton, E. (2013). *Robust variance estimation with dependent effect sizes: Practical considerations and a software tutorial in Stata and SPSS.* Research Synthesis Methods, in press. 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

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.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013, February 4). A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games. *Journal of Educational Psychology, 105*(2), 249-265.

SRI Education

BILL & MELINDA GATES *foundation*