Simulations for STEM Learning: Systematic Review and Meta-Analysis

Executive Summary

DRAFT

May 2013

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

This report is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

Suggested citation:
D'Angelo, C., Rutstein, D., Harris, C., Bernard, R., Borokhovski, E., Haertel, G. (2013). *Simulations for STEM Learning: Systematic Review and Meta-Analysis (Executive Summary).* Menlo Park, CA: SRI International.

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
Phone: 650.859.2000

www.sri.com

Simulations for STEM Learning: Systematic Review and Meta-Analysis

Executive Summary

Prepared By:

Cynthia D'Angelo
Daisy Rutstein
Christopher Harris
Robert Bernard
Evgueni Borokhovski
Geneva Haertel,

DRAFT

May 2013

Contents

Executive Summary

Overview

This executive summary presents an overview of the process and initial findings of a systematic review and meta-analysis of the literature on computer simulations for K–12 science, technology, engineering, and mathematics (STEM) learning topics. Both quantitative and qualitative research studies on the effects of simulation in STEM were reviewed. Those that reported effect size measures or the data to calculate effect sizes were included in the meta-analysis. Important moderating factors related to simulation design, assessment, implementation, and study quality were coded, categorized, and analyzed for all the articles.

Two research questions guided the review and meta-analysis:

1. What is the difference in outcome measures between K–12 students who receive simulations as a form of instruction and K–12 students who receive some other kind of instructional treatment?
2. What is the difference in outcome measures between K–12 students who receive simulations that are supplemented or modified with some other form of instructional treatment (e.g., simulation plus scaffolding) and simulations without modifications?

Highlighted here are the important preliminary findings of the study thus far.

Background

With the rise in computing and lowering of computer costs has been an increase in the use of simulations. A simulation, for the purposes of this study, is a computer-based interactive environment with an underlying model. In the STEM field in particular, real equipment can be difficult to obtain, so simulations enable students to experience phenomena they normally would not be able to experience firsthand. For example, simulations can take the place of laboratory equipment that might be too expensive or dangerous to have in a school. Simulations can also be used to explore phenomena that occur over long or extremely short time periods in a way that can easily fit into a class period. With simulations, students can also manipulate variables and see the results of multiple experiments without having to actually replicate them. (See Quellmalz & Pellegrino, 2009, for a review of the use of simulations in K–12 settings and the affordances of simulations that can affect student outcomes.)

In view of these benefits, it is believed that using simulations in the classroom can help improve learning. Several literature reviews (e.g., Scalise et al., 2011; Smetana & Bell, 2012) have examined whether and how simulations aid the improvement of student learning. However, this literature has not been quantitatively and systematically analyzed to determine whether simulations do in fact have an effect on student learning.

In the summer of 2012, the Bill & Melinda Gates Foundation, in cooperation with the MacArthur Foundation, made a significant investment to establish and support the Games Assessment and Innovation Lab (GlassLab), which includes top game developers, assessment experts, and researchers from multiple fields. The goal of GlassLab is to transform learning and formative assessment through digital games. During the planning stages of the investment, the program was divided into two teams — an investment in a program team (GLASSLab) and a second investment in a research team (GLASSLab-Research) — to mitigate conflict of interest and guarantee independent validation of assessments developed by the program. It was determined by all those associated with GlassLab that the GlassLab development team would design and develop state-of-the-art game-based assessments. Independently, GlassLab-Research would conduct research on the qualities, features, validity, reliability, and effectiveness of the games and assessments that are embedded within the gaming environments produced by GlassLab. The meta-analysis and systematic review of the simulation literature described in this report is part of the larger GlassLab-Research project.

Defining a Simulation

The first goals of this project were to develop a working definition of simulation and to determine how simulations differ from other computer-based learning tools. The research team recognized that a continuum exists, with basic computer-based visualizations or animations at one end and complex video games at the other. We focused solely on the middle area, computer-based simulations that are neither simple visualizations nor involved games. To define this continuum further, the team made two important distinctions.

The first was to differentiate a simulation from a game. [1] We defined a game as having clear goal states and a built-in reward system (such as points or currency) tied to these goal states. For the meta-analysis, a computer-based tool was classified as a game if the user needed to reach levels or achievements in order to progress. Comparatively, a simulation was something that allowed users to be more focused on a specific phenomenon or activity than on achieving non-learning-based goals.

The other distinction was between a simulation and a visualization. This distinction hinges on the important concept of interaction with a scientific model. Simulations, as defined here, must be constructed with an underlying model that is based on some real-world behavior or natural/scientific phenomena (such as models of the ecosystem or simulated animal dissections). The important criterion is that the simulation include some interactivity on the part of the user, centered on inputs and outputs of the model. Otherwise, the tool was labeled as a visualization rather than a simulation.

---

1 Another research group is performing a meta-analysis on games for learning (Clark, Tanner-Smith, Killingsworth, & Bellamy, 2013) as part of the larger GLASSLab-Research project. The game/simulation boundary resulted from a discussion between this group and our team to ensure little overlap or gap existed between our searches. For example, virtual worlds fell along the boundary between simulations and games, and the two groups decided that they should be part of the simulation meta-analysis.

Other Extant Literature Reviews

Reviews exist of simulations or computer-based tools that help students learn various STEM concepts. Some of them are focused on a very narrow range of simulation studies or on overall trends of the findings of these studies, but none conducted a comprehensive quantitative meta-analysis. For example, in a recent review Smetana and Bell (2012) looked at computer simulations that are meant to support science instruction and learning. They found that most (49 of 61) studies showed positive impacts of the use of simulations. Although the studies discussed are thoroughly explained and categorized, the search procedures were not very well documented, and our research team identified many key researchers and articles as missing from the review.

Another recent review (Scalise et al., 2011) also examined learning through science simulations. This review was on software for grades 6–12, particularly virtual laboratory simulations. Another review (Clark, Nelson, Sengupta, & D'Angelo, 2009) looked at science learning gains from both simulations and games. This paper mostly described available simulations/games and overall findings from studies and reported details in a few select areas.

None of these reviews were proper meta-analyses where effect sizes across a series of studies were calculated and compared. The study described in this report includes a meta-analysis and is building on these previous reviews while taking on additional challenges. We are examining not only the effectiveness of simulations for STEM learning, but also the features of simulations that contribute to learning gains, the types of research and study designs that are most effective for determining these gains, any moderating variables that influence learning gains, and details of the assessments and measures used to determine learning gains. Some of these factors are included in a formal quantitative meta-analysis (described in this summary) whereas others are the subject of a more detailed systematic qualitative description and review (forthcoming in the third quarter of 2013).

The preliminary study results presented here provide a look at the factors that influence learning science and engineering in computer-based simulations. The final report will include more details on these factors as well as implications for how to design and build these simulations and how to assess learning in these environments.

Meta-Analysis

A meta-analysis is the systematic synthesis of quantitative results from a collection of studies on a given topic (Borenstein, Hedges, Higgins, & Rothstein, 2009). Many terms have been used to describe literature reviews, such as research synthesis, research reviews, and narrative reviews (Cooper, 2010). While some of these terms are used interchangeably with meta-analysis (Cooper favors *research synthesis*), what sets a meta-analysis apart from other literature reviews is the quantitative and systematic nature of the data collection and analysis.

Part of the systematic approach in a meta-analysis is to document the decisions that are being made about the collection of the articles and the steps of the analysis. This allows for the study to be replicated. The approach also calls for the specification of the research questions guiding the analysis because two researchers examining the same set of articles may be asking different questions and thus may arrive at different results. Another part of being systematic in the approach is to help ensure that articles are collected and reviewed in a carefully organized manner to make sure the study is as inclusive as possible (Borenstein et al., 2009). In a meta-analysis articles are included based on pre-defined criteria and not because of 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, and 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 found through various inferential statistics, such as *t*-tests or ANOVAs. For example, a study 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.

The magnitudes of effect sizes can be categorized into different groups. For Cohen (1988), one way to think about categorizing effect sizes was that small effect sizes (.2 to .3) are those that are barely detectable by the naked eye, medium effect sizes (.4 to .6) are those that can be detected visually, and large effect sizes (greater than .7) are those that could not be missed by a casual observer. It is important to remember that effect sizes are dependent not just on the mean difference between two groups, but also the standard deviation of those groups. For example, there is an average height difference between 15- and 16-year old girls, but there is a lot of variation within each of those age groups, so this would correspond to a relatively small effect size. However, when comparing 13- and 18- year old girls, there is a much larger average height difference, and even with a similar amount of variation within each age group, this would correspond to a larger effect size.

In addition, if the effect size is consistent across a collection of articles, then an overall effect size can be estimated that is both robust and applicable to the type of studies used (Borenstein et al., 2009). Further exploration of effects using moderating variables can be performed to understand what particular variables contribute to the results.

The tools of meta-analysis enable researchers to look across a large number of similar studies to determine whether certain kinds of interventions have consistent effects. This is a powerful kind of analysis that, when combined with the systematic nature of a meta-analytic review, presents a solid view of the current state of research and findings in a field.

Methods

Scope

This meta-analysis is concerned with the effectiveness of computer simulations used in instructional settings. The scope was limited to simulations in STEM contexts or content in order to align with the GLASSLab game developers' objectives. It was also decided to limit analysis to only studies with participants in the K–12 grade range (although simulations did not need to occur in a formal school setting). The results will therefore be applicable directly to simulation and curriculum designers working in these grade levels. The list of possible outcome measures was kept broad at this point to be responsive to what was in the literature.

Initial Search

The research team used three well-known and comprehensive databases to ensure the search covered all the relevant literature and journals: the Education Resources Information Center (ERIC) (http://www.eric.ed.gov/), PsycINFO (http://www.apa.org/psycinfo/), and Scopus (http://www.scopus.com/). From discussions with a research librarian, we determined that because of the overlapping coverage and journal availability, these databases should be able to capture nearly all the relevant literature on learning simulations.

To identify as many articles as possible, we performed the searches using the title, abstract and keyword or descriptor fields in the databases. We decided to keep the search terms relatively broad in order to capture a large number of potential articles but not too broad. Specifically, we used the combination of the terms *simulation* or *computer simulation* along with STEM content terms such as *science education* and *mathematics education*. Searching for *simulation* alone would have produced an order of magnitude more articles than the search we ended up with. That volume of articles would have taken a prohibitively long time to properly sort through, given our resource constraints.

The initial search terms included the STEM domains (*science, technology, engineering*, and *mathematics* and their subtopics, such as *biology* and *chemistry*) and *simulation* or *computer simulation* as primary search terms. Other topics, such as *21st century skills* were included in coding, study categorization, and analysis. For example, a study about problem solving in the context of science learning would be included in the search because of the emphasis on science learning and because the simulation features, assessments, and results relating to problem solving are reported along with other science content-related features, assessments, and results.

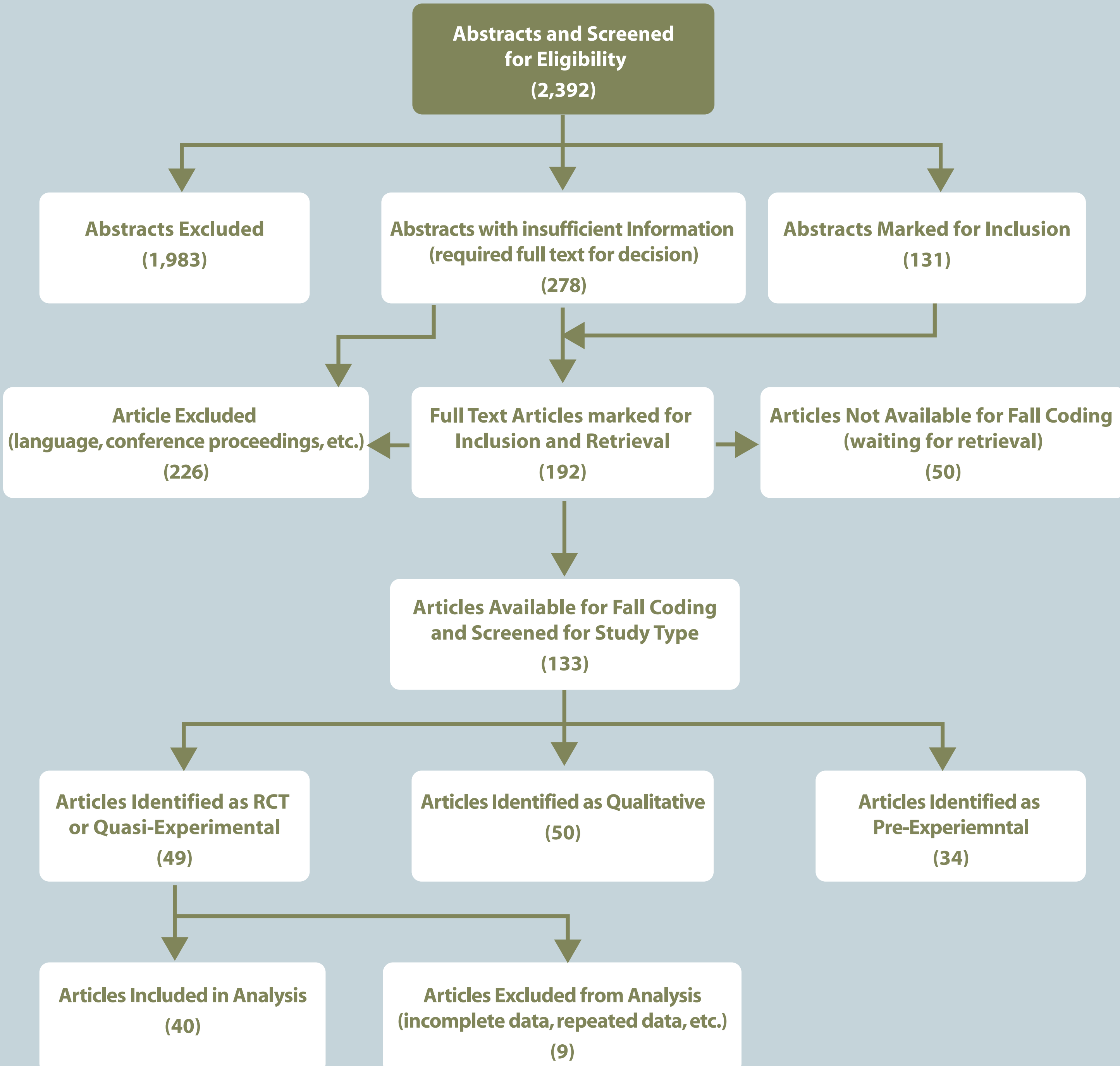
Only articles published between 1991 and 2012 (inclusive) were included in the study. The majority of simulation-based education research studies were conducted during this time, and any studies done before 1991 are likely to concern technologies that are out of date and would not be helpful to contemporary researchers, educators, and designers. Only peer-reviewed journals were included, and only articles in those journals (i.e., not editorials). The decision to exclude literature such as conference proceedings and non-peer-reviewed articles was to ensure a high quality of research and keep the pool of articles manageable. Additionally, to be included in the quantitative meta-analysis portion, studies needed to include the relevant quantitative information needed for the effect size calculations.

Method Overview

Exhibit 1 presents an overview of the search and coding process. Overall, 2,392 abstracts were reviewed, resulting in full-text retrieval of about 200 primary research studies potentially suitable for the analysis. Through a thorough review of full-text documents, 133 studies were retained for further analysis. Of these, 49 were determined to be research articles including either an experimental or quasi-experimental design. Of those, 9 were determined to contain incomplete or repeated data and were excluded from our analysis. The remaining 40 studies yielded 104 effect sizes, 67 of which were in the achievement outcome category, 11 were in the attitudes category, and the remaining 26 that fell into other categories (such as inquiry skills).

The sections that follow describe the methods at each stage in this process.

Exhibit 1. Abstract and Article Retrieval and Coding Process



Abstract Screening Stage

The abstracts for the 2,392 articles produced from the initial search of the databases were collected using the citation management program Mendeley[2]. The simulation meta-analysis team developed an exclusion coding scheme, with two team members per article coding each abstract. Articles coded for exclusion were assigned to one or more exclusion categories (Exhibit 2). Our search strategy was to find a large number of articles that met certain criteria (e.g., year of publication, source) and then exclude individual articles that did not meet our other criteria (e.g., research study, interactive simulation) for one or more reasons. These exclusion categories further defined our search parameters and inclusion criteria.

Specifically, we wanted to look at studies that involved students in kindergarten through high school, regardless of whether the study took place within a formal learning environment. Thus, studies involving students outside the K–12 grade range were excluded at the abstract screening stage. Because we also needed to check whether the simulation described in the study met our definition of simulation, many of the exclusion categories dealt with this (e.g., not computer based, visualization, game). We also excluded articles that did not describe a research study.[3] Many articles contained descriptive information about a simulation but did not present any data or evidence that an investigation had been performed, so these were excluded for not being research based.

High agreement existed among the coders, with the pairs agreeing on 86.1% of the first-round abstract coding. Most of the disagreements (66.1%) occurred when coders could not agree on the exclusion category or categories. Two researchers resolved all the disagreements by reviewing the abstracts and discussing the disagreements.

From the review of the abstracts, 131 (5%) of the original articles were determined to match all our inclusion criteria and appeared to address one or both of the research questions. For about 300 (12%) of the articles, information in the abstract alone was insufficient for making a decision. Full texts of those articles were obtained, and two researchers coded them using the same codes as for the abstracts. The remaining 83% of the articles were excluded for one or more reasons.

2 http://www.mendeley.com

3 Qualitative research methods were included at this stage, although the outcomes associated with these methods (such as student interviews) were not analyzed for this report.

Exhibit 2. Abstract Screening Results: Exclusions

Exclusion reason	Number of Abstracts	Percentage of Abstracts
Not K–12 grade range	946	39.5
Not a research-based article	882	36.9
Simulation is not part of instruction	439	18.4
Does not fit our simulation definition (other)	293	12.2
Review or trend article	119	5.0
Content is not STEM related	119	5.0
Not computer based	96	4.0
Game	63	2.6
Visualization	25	1.0

Note: Abstracts could be coded for more than one exclusion reason.

Article Screening Stage

Once the abstracts were screened, we collected complete texts of all the included articles. Simultaneously, the research team developed and refined a coding scheme for them. The coding scheme captures information about the research questions, the research design, the study variables, the effect size data, the assessments used, the features of the simulations, implementation information, and participant information.

Two members of the team went through the list of the full texts, read them, and identified which articles were quasi-experimental or randomized controlled trials and had enough data to be included in the meta-analysis. Inter-rater agreement for this full-text manuscript inclusion/exclusion was 94.50% (κ= 0.89). Each of these articles was coded by two team members on a subset of the codes that were deemed most relevant to the meta-analysis at this time. (The list of the articles included at this stage in the study is in Appendix A in the full report.)

Codebook Development

The research team developed a set of codes to describe the studies (e.g., demographics, methodological study features) and their substantive characteristics for use in subsequent moderator variable analysis. This was an iterative process that entailed identifying an initial set of codes with a subset of the articles and then refining and creating new codes as the review of articles proceeded. The initial codes described features of the articles that we wished to capture. All the articles were coded with the finalized coding scheme. Some of the codes were applied at the article or study level (pertaining to research design or location of the study), whereas others were applied at the effect size level (pertaining to specific comparisons and findings of the studies). The codes fell into six broad categories:

1. Demographic information (location of study, ages of participants, language of instruction)
2. Study information (research question, STEM topic)
3. Methodological information (research design, group equivalency, attrition)
4. Assessment information (source of assessment, type of measures)
5. Simulation information (type, collaboration, flexibility, platform)
6. Implementation information (setting, curriculum, time/duration/frequency).

The entire codebook with detailed descriptions of each code used and its value options is in Appendix B in the full report. A sample page from the FileMaker database created for coding is in Appendix C in the full report.

Quantification in Meta-Analysis

The basic metric and unit of analysis in a meta-analysis is an effect size. The one used in this meta-analysis is a *d*-type effect that expresses the standardized difference between the means of two groups. Cohen's *d* (Cohen, 1988) has become the more accepted form of the *d*-type effect size. Cohen's *d* is calculated by pooling the standard deviations of the experimental and control groups and using this new standard deviation as the divisor of the mean difference.

In addition, Hedges & Olkin (1985) introduced a multiplier to Cohen's *d* that corrects for small-sample bias. This adaptation is generally referred to as Hedges' *g*. The effect sizes of small samples (generally around 40 participants) are adjusted downward slightly, while larger samples remain unaffected. As a result, most reviewers convert all *d*-type effect sizes to Hedges' g because it corrects bias in small sample studies without affecting larger samples.

Synthesizing Effect Sizes

Effect sizes are always weighted at the synthesis phase, where effect sizes are combined into an overall average. There are multiple models to consider at this stage: fixed-effect, random-effects, and a mixed-effect model. The weights for the fixed-effect model[4] and the random-effects model are different, owing to the theoretical definitions of the models (e.g., Borenstein, Hedges, Higgins & Rothstein, 2010). We will use the fixed-effect model to estimate heterogeneity of *k* effect sizes (where *k* indicates the number of effect sizes in the synthesis) and the random-effects model to estimate the weighted average effect size ($g+$) and the 95th confidence interval within which the mean resides.

Fixed-effect model. The underlying assumption of the fixed-effect model, where effect sizes are weighted by their inverse variance (i.e., $W_{g(Fixed)} = \frac{1}{V_g}$), is that a precise fixed average effect size can represent all studies in the meta-analysis that are essentially alike in terms of research design, treatment definition, outcome measures, and sample demographics. There are two primary outcomes of a first-level synthesis of a distribution of *k* effect sizes under the fixed-effect model: (1) the average weighted effect size of k effect sizes ($g+$ is the statistical symbol for the weighted average) and associated statistics (i.e., standard error, variance, the upper and lower limits of the 95th confidence interval, a z-test and associated probability) and (2) heterogeneity assessment and its associated test statistics. For heterogeneity analysis, a *Q*-statistic (Cochran's *Q*) is created from the squared sum of each effect size subtracted from the average effect size. The *Q*-statistic is a sum of squares that is assessed using the chi-squared distribution with $p - 1$ degrees of freedom. Failure to reject the null hypothesis leads to the conclusion that the distribution is homogeneous (i.e., between-study variability does not exceed chance expectations). A significant *Q*-value denotes heterogeneity that exceeds the expected level of chance. Higgins and colleagues (Higgins, Idson, Freitas, Spiegel, & Molden, 2003) developed I^2 as a more intuitive measure of heterogeneity. I^2 ranges from 0.0 to 1.0 and is read as a percentage of between-study variability contained in total variability.

Random-effects model. The random-effects model is considered most appropriate when studies in the meta-analysis differ in terms of methodology, treatment definition, demographics, and the like. The inverse variance weights include the between-study variance term τ^2 (i.e. ,$W_{g(Random)} = \frac{1}{V + \tau_g^2}$).Studies are not assumed to be alike except in the sense that they all address the same general research question (e.g., the effects of educational simulations on learning). Each study is deemed to be a random sample from a micropopulation of like studies. There is no heterogeneity assessment since all between-study variability is resolved within each study.

Mixed-effect model. Moderator variable analysis involves comparisons between/among levels of coded study features and is considered a secondary level of comparison. The mixed-effects model is, as the name implies, a combination of the characteristics of the fixed and random models. Average effects at each level of the moderator variable are synthesized using the random-effects model with τ^2 calculated separately for each level. Synthesis across levels is performed using the fixed-effect model.

4 For a full description of these models and their underlying assumptions, see Hedges & Olkin (1985), Borenstein, Hedges, Higgins and Rothstein (2009), and Pigott (2012),

Results

Research Questions

We identified three outcome measure categories in our review of the literature on computer simulations in K–12 STEM education: achievement measures, attitude measures (both content area related and technology related), and inquiry and reasoning skills. Two research questions were found in the literature of educational simulations related to the three outcome measures:

1. What is the difference in outcome measures between K–12 students who receive simulations as a form of instruction and K–12 students who receive some other kind of instructional treatment?
2. What is the difference in the outcome measures between K–12 students who receive simulations that are supplemented or modified with some other form of instructional treatment (e.g., simulation plus scaffolding) and simulations alone?

Descriptive Results

The 40 articles selected for inclusion in the meta-analysis were coded using the definitions and rules in the codebook. Some of the variables were included as moderator variables, as described below. Others were coded to use as descriptive variables to help us better understand the pool of articles selected for study (including demographics of participants, specific simulation topic, etc.). Exhibits 3 (study level) and 4 (effect size level) detail the results of coding for the pertinent descriptive variables.

Exhibit 3. Descriptive Results at Study Level (40 Studies)

Variable	Frequency
Location of study	
North America	13
Europe	17
Asia	5
Not indicated	3
STEM domain	
Science	33
Mathematics	4
Engineering	2
Technology	1
Grade level of participants	
Kindergarten– grade 5	4
Grades 6–8	12
Grades 9–12	23
Multiple ranges	1

The 40 studies were found to contain 104 effect sizes for the purposes of this meta-analysis. Each effect size represents a study's comparison that falls under one of the two research questions and one of the outcome measures. A single article could have multiple effect sizes if it reported multiple outcomes for a single control/treatment comparison or if multiple groups were being compared on a single outcome (e.g., in a factorial design). Inter-rater agreement for effect size identification and calculation (i.e., accuracy of data extraction and selection and application of equations) was 95.50% ($\kappa = 0.97$).

Exhibit 4. Descriptive Results at Effect Size Level (104 Effect Sizes)

Variable	Research Question	
	1: Simulation vs. No Simulation ($k = 59$)	2: Simulation Plus Enhancement vs. Simulation Alone ($k = 45$)
Outcome measures		
Achievement	36	31
Attitude	10	1
Other (reasoning, inquiry, etc.)	13	13
Simulation type		
Phenomenon simulation	16	23
Virtual lab	15	18
Agent based	4	0
Virtual world	0	0
Other	3	4
Not Indicated	21	0
Assessment delivery mode		
Embedded in simulation	1	7
Tech based but not embedded	1	7
Not tech based	53	17
Not indicated	4	14
Assessment source		
Researcher designed	45	39
Teacher designed	5	1
District, state, or national test	4	0
Other standardized test	0	4
Curriculum test	0	0
Not Indicated	5	1

Meta-Analysis Results

Publication Bias and Outlier Analysis

We investigated potential publication bias by inspecting funnel plots (i.e., effect size by standard error) and through the use of statistical tools that are resident in *Comprehensive Meta-Analysis™* (*CMA*, Version, 2.2057, Borenstein et al., 2010). We found no serious publication bias in achievement outcomes (i.e., potential studies that were not located). For attitude outcomes, however, the distribution appeared bimodal, suggesting that possibly attitudes toward instruction and attitudes toward content were mixed in this collection. Nonetheless, the number of effect sizes (k = 11) was deemed too small to bifurcate the collection.

We used the "one study removed procedure" resident in CMA to investigate potentially outlying effect sizes. As a result, one very large effect size (g = +4.57) was reduced to the next highest effect size of 2.10.

Methodological Quality

Seven methodological characteristics were identified, coded, and tested to determine whether the collection of studies contained systematic bias due to the methods the primary researchers used that might alter the interpretation of results. In the cases of the variables with asterisks in the following list, the number of studies was reduced from k = 67 because of missing data (see variables below for final numbers):

- **Research design** – Randomized controlled trials ($g+$ = 0.59, k = 14) vs. quasi-experimental designs ($g+$ = 0.47, k = 53), $Q_{Between}$ = 0.42, df = 1, p = .52, *NSD*[5]
- **Methods of effect size extraction** – Exact descriptive statistics ($g+$ = 0.66, k = 29) vs. estimated from inferential statistics ($g+$ = 0.35, k = 34) vs. reported by researcher ($g+$ = 0.69, k = 4), $Q_{Between}$ = 6.76, df = 2, p = .03, *SD*
- **Instructor equivalence*** – Same instructor ($g+$ = 0.74, k = 14) vs. different instructor ($g+$ = 0.55, k = 25), $Q_{Between}$ = 0.85, df = 1, p = .36, *NSD*
- **Material equivalence*** – Same materials ($g+$ = 0.47, k = 42) vs. different materials ($g+$ = 0.40, k = 17), $Q_{Between}$ = 0.23, df = 1, p = .63, *NSD*
- **Time-on-task (TOT) equivalence*** – Same TOT ($g+$ = 0.46, k = 46) vs. different TOT ($g+$ = 0.39, k = 10), $Q_{Between}$ = 0.15, df = 1, p = .70, *NSD*
- **Source of outcome measure*** – Standardized state, national or district ($g+$ = 0.74, k = 4) vs. other standardized ($g+$ = 0.67, k = 4) vs. researcher made ($g+$ = 0.44, k = 49) vs. teacher made ($g+$ = 0.57, k = 4), $Q_{Between}$ = 4.73, df = 3, p = .19, *NSD*.

Except for *methods of effect size extraction*, all the methodological moderator variables were uniform across levels. We concluded that there was no severe bias due to the research practices in the collection.

5 *NSD* = No significant difference at the $\alpha = 0.05$ level.

Overall Effects by Category of Outcome Measure and Research Question

Exhibit 5 presents average effect sizes for the outcome categories of achievement, attitude, and scientific inquiry and reasoning skills separately for each research question. All average effects were positive and significantly different from zero. The random-effects model analyses of achievement outcomes produced an average effect $g+ = 0.67$ ($k = 36$) for Research Question 1, suggesting a moderate to strong (Cohen, 1988) positive influence of simulation-based instructional interventions compared with non-simulation-based methods of instruction (17% increase of the average treatment over the control group). This average effect was higher than the average effect on achievement outcomes ($g+ = 0.31$, $k = 31$) of enhanced simulations compared with simulations alone (Research Question 2). Fixed-model analyses showed that both average effects were significantly heterogeneous ($Q_T = 98.47$ and $Q_T = 129.18$, respectively, $p < .001$), warranting further exploration of variability through moderator variable analyses.

Attitude data for Research Question 1 produced an average effect size of $g+ = 0.87$ ($k = 10$) that is significantly heterogeneous ($Q_T = 141.07$, $p < .001$). The single effect size for Research Question 1 in this outcome category ($g = -1.22$) did not allow for further interpretation. Also, we did not conduct heterogeneity analyses or subsequent moderator variable analyses on effect sizes in the outcome

Exhibit 5. Overall Effects by Category of Outcome Measure and Research Question

Variable	*k*	*g+*	Lower 95th	Upper 95th
Achievement				
Question 1: Simulation vs. no simulation	36	0.67	0.51	0.82
Heterogeneity (fixed-effect model)	$Q_T = 98.47$	$df = 35$	$p < .001$	$I^2 = 64.46$
Question 2: Simulation plus enhancement vs. simulation alone	31	0.31	0.12	0.49
Heterogeneity (fixed-effect model)	$Q_T = 129.18$	$df = 30$	$p < .001$	$I^2 = 76.78$
Attitudes				
Question 1: Simulation vs. no simulation	10	0.87	0.12	1.61
Heterogeneity (fixed-effect model)	$Q_T = 141.07$	$df = 9$	$p < .001$	$I^2 = 93.62$
Question 2: Simulation plus enhancement vs. simulation alone	1	−1.22 (Not part of the comparison)		
Heterogeneity (fixed-effect model)	N/A	N/A	N/A	N/A
Inquiry and reasoning skills				
Question 1: Simulation vs. no simulation	5 (all from 2 studies)	0.20	-0.15	+0.56
Heterogeneity (fixed-effect model)	QT = 7.47	$df = 4$	$p = .11$	$I^2 = 46.44$
Question 2: Simulation plus enhancement vs. simulation alone	4 (all from 1 study)	0.60	0.35	0.85
Heterogeneity (fixed-effect model)	$Q_T = 3.07$	$df = 3$	$p = .38$	$I^2 = 2.36$

category of scientific inquiry and reasoning skills for either research question because all nine effect sizes were derived from only three primary studies (from a single study for Research Question 2). However, the average effect sizes in this outcome category were positive (i.e., $g+ = 0.20$, $k = 5$ for Research Question 1 and $g+ = 0.60$, $k = 4$, for Research Question 2, with only the latter being statistically significant).

Forest plots of the Hedges' *g* effect size, confidence interval, and weight of each of the included studies for the research questions (both achievement and attitude outcomes) are in the full report.

Demographic and Instructional Moderator Variable Analysis

Achievement outcomes for Question 1. The analyses of demographic moderator variables for Research Question 1 showed that effects were not significantly different across grade levels ($Q_B = 0.11, p > .05$), although effect sizes for the STEM content area domain varied significantly ($Q_B = 13.42$, $p < .001$). This was particularly true for the science content area, producing a strong significant effect of $g+ = 0.67$, $k = 33$. The effect of simulation use in teaching mathematics was not significantly different from zero. This finding should be viewed with extreme caution, however, because the mathematics data were based on only two effect sizes. Due to these low numbers, additional searching with alternate keywords (such as "linked representations") was done and more mathematics articles will be part of the final analysis and report.

The findings for instructional moderator variables (Exhibit 6) portrayed particular characteristics of the experimental intervention only. None of these analyses revealed significant differences among effect sizes at different levels of the instructional moderator variables.

Exhibit 6. Instructional Variables Related to the Simulation Treatment for Question 1

Variable	*k*	*g+*	Lower 95th	Upper 95th	Q_B *(df, p)* Conclusion
Simulation Type*					
Virtual lab	9	0.85	0.53	1.18	
Phenomenon simulation	15	0.57	0.34	0.79	
Agent based	4	0.74	0.50	0.98	
***Q*-between**					2.29 (2, .32), *NSD*
Group work*					
Individual	8	0.78	0.56	0.99	
Dyads	10	0.53	0.15	0.91	
Small groups (> 2)	9	0.88	0.52	1.24	
***Q*-between**					1.86 (2, .39), *NSD*
Post hoc: Individual vs. dyads					1.25 (1, .26), *NSD*
Post hoc: Dyads vs. small groups					1.73 (1, .19), *NSD*
Assessment delivery model*					
Not by technology	33	0.68	0.51	0.84	
Technology, but not embedded	1	0.65 Not part of the comparison			
***Q*-between**					N/A

* "Not reported" and "Other" data removed.

Achievement outcomes for Question 2. The effects of simulation plus enhancement across grade levels were not significantly different ($Q_B = 2.68, p > .05$), but within STEM domains the average effect for instruction in mathematics was negative and significantly different from average effect sizes for science and engineering (Exhibit 7). Also, average effect sizes did not vary significantly across any level of the instructional moderator variables (Exhibit 8).

Exhibit 7. Demographic Moderator Variables Related to the Simulation Plus Treatment Condition for Question 2

Variable	*k*	*g+*	Lower 95th	Upper 95th	tt
Grade range					
Kindergarten–grade 5	2	0.31	0.01	0.62	
Grades 6–8	10	0.43	0.09	0.76	
Grades 9–12	17	0.18	-0.09	0.45	
Multiple ranges	2	0.75	-0.01	1.52	
***Q*-between**					2.68 (3, .44), *NSD*
STEM domain					
Science	22	0.43	0.24	0.63	
Math	7	-0.33	-0.76	0.09	
Engineering	2	0.75	-0.01	1.52	
***Q*-between**					11.52 (2, .003), *SD*

Exhibit 8. Instructional Moderator Variables Related to the Simulation Plus Treatment for Question 2

Variable	*k*	*g+*	Lower 95th*	Upper 95th*	Q_B *(df, p)* Conclusion
Simulation type*					
Virtual lab	13	0.53	0.30	0.76	
Phenomenon simulation	15	0.21	-0.03	0.46	
***Q*-between**					3.36 (1, .07), *NSD*
Group work*					
Individual	17	0.19	-0.06	0.44	
Dyads	5	0.36	-0.08	0.79	
Small groups (> 2)	5	0.43	-0.19	1.05	
***Q*-between**					0.79 (2, .67), *NSD*
Assessment delivery model*					
Not by technology	17	0.32	0.06	0.58	
Technology, but not embedded	5	0.53	0.15	0.90	
***Q*-between**					0.79 (1, .38), *NSD*

* "Not reported" and "Other" data removed.

Discussion of Current Results

These initial findings indicate that simulations have promise for improving students' learning outcomes in STEM topics. Although further analysis is required both with these studies and the qualitative and pre-experimental studies identified in the literature search, many high- level findings can be discussed.

The between-study variability across all outcomes and research questions tended to exceed what would be expected by chance sampling. This suggests that to perform appropriate analyses on the effects of simulations on learning, separating the different outcome measures and research questions was necessary. The results are presented at this level. Conflating multiple research questions or outcomes would lead to inappropriate conclusions. Although the average effect sizes were positive in each of the groupings, the nature of the effects was slightly different for the different types of studies and outcomes and therefore should not be directly compared.

Simulation treatments were shown to have an advantage in learning achievement over non-simulation instruction. Many prior literature reviews had reached a similar conclusion, so this is not a surprise. However, this meta-analysis was able to quantify the magnitude of average improvement due to simulations and look at specific moderator variables. The results showed that no significant differences existed across the K–12 age groups included in this study. Nor were significant differences found across different types of simulations or across group size (individual vs. dyads vs. small groups). It seems clear that simulations, in many different configurations or contexts within the classroom, do improve student science learning compared with not using simulations. Other STEM disciplines were underrepresented in this study because of the lack of articles meeting our criteria in engineering,[6] technology, and mathematics.

We also found advantages of simulation treatment in improving student attitudes using simulations. Because this analysis combined different types of attitudes (content related and technology related), however, the small number of quantitative studies including attitude measures precludes drawing many conclusions based on this result.

Simulations supplemented or modified with some other form of instructional treatment (e.g., simulation plus scaffolding) provided modest improvements in learning achievement over simulations alone. Many different modifications or enhancements were used in the 31 studies analyzed. The types of modifications did cluster in a few general areas, specifically, scaffolding, representations, haptic feedback (feedback involving touch), and cooperative learning. A preliminary finding based on these categories of modifications showed that scaffolding within a simulation had a strong positive effect, as did cooperative learning, whereas representations had a more mixed overall effect. Further analysis of these studies will be conducted in 2013. We found no significant differences across age groups, simulation type, or group size. There was, however, a significant difference across STEM domains, with mathematics simulations having a much lower *g+* than either science or engineering simulations. A comparison based on attitude outcomes for simulations supplemented or modified with some other instructional treatment was not possible because there was only one effect size in this category.

This study had limitations, some of which will be addressed in the coming months of work. One concerned certain moderator variables and lack of coder agreement.

6 There were many studies involving engineering education, but they were excluded because they used college-age or older students. A follow-up project could examine studies with these older students if engineering education is an important area of interest.

The variables relating to the length and duration of the treatments mentioned in the studies were particularly difficult for researchers to code reliably. There is no standard way of reporting this type of data, and the researchers had to make many judgments that did not always align. As this seems to be an important variable to examine, we will address this as this analysis continues.

Other variables of interest are those relating to specific features of the simulations in these studies. As we read through and coded the abstracts and articles, we noted that not as much information was given about the simulations as would be required to code for certain simulation features as moderator variables. We noticed that articles that did contain this type of information typically did not also include any kind of research or study and were therefore excluded from the systematic review at either the abstract or article stage. Articles with study details and outcome measures typically did not have as many details about the design or features of the simulation. Additionally, articles on some simulations that were commonly used (such as NetLogo) might not contain as many details because they are taken as known. One of our tasks for the coming months is to try to follow up with study authors either directly or through other articles to seek more information about the specific features of these simulations.

The lack of mathematics, technology, and engineering simulation studies was surprising. The engineering problem can be solved by simply including studies involving college-age students. Current work underway includes coding and analyzing a new batch of articles on mathematics simulations that were found using alternate keywords (e.g., *linked representations* instead of *simulation*).

As this analysis continues, more information will be uncovered about how simulations can improve student learning as well as more details about what instructional contexts and features are most beneficial to students.

References

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

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111.

Clark, D. B., Nelson, B., Sengupta, P., & D'Angelo, C. M. (2009). Rethinking science learning through digital games and simulations: Genres, examples, and evidence. Invited Topic Paper in the *Proceedings of The National Academies Board on Science Education Workshop on Learning Science: Computer Games, Simulations, and Education.* Washington, D.C.

Clark, D. B., Tanner-Smith, E. E., Killingsworth, S., & Bellamy, S. (2013). *Digital games for learning: A systematic review and meta-analysis* (Vol. 1). Nashville, TN: Vanderbilt University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks, CA: Sage.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Higgins, E. T., Idson, L. C., Freitas, A. L., Spiegel, S., & Molden, D. C. (2003). Transfer of value from fit. *Journal of Personality and Social Psychology, 84,* 1140–1153.

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *Practical guide to modern methods of meta-analysis. National Science Teachers Association.* [ERIC Document Reproduction Service No. ED 309 952].

Pigott, T. (2012). *Advances in meta-analysis (statistics for social and behavioral sciences)*. New York: Springer.

Quellmalz, E. S., & Pellegrino, W. (2009). *Technology and testing. Science Magazine*, 323, 75–79.

Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K., & Irvin, P. S. (2011). Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching, 48*(9), 1050–1078.

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education, 34*(9), 1337–1370.

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
Phone: 650.859.2000

www.sri.com

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

BILL & MELINDA GATES foundation