Simulations for STEM Learning: Systematic Review and Meta-Analysis

SRI Education

March 2014

Developed by SRI Education 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. (2014). *Simulations for STEM Learning: Systematic Review and Meta-Analysis.* Menlo Park, CA: SRI International.

SRI Education

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

www.sri.com/education

Simulations for STEM Learning: Systematic Review and Meta-Analysis

Prepared By:

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

SRI Education

March 2014

Contents

Overview

This report 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. Studies 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.

This review and meta-analysis is focused on two research questions:

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 and those who receive the same simulations that are modified with some form of instructional enhancement, such as dynamic representations, meta-cognitive support, or extended feedback?

This report includes a full reporting of the research activities of this project and all of the results and findings.

Background

With the rise in computing and reduction of computer costs, the use of simulations has increased. 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 widely held 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 conflicts of interest and guarantee independent validation of assessments developed by the program. The GlassLab program team (GlassLab) was tasked to design and develop state-of-the-art game-based assessments. Independently, the research team (GlassLab-Research) would conduct research on the qualities, features, inferential 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. In our definition, a computer simulation is a tool used to explore a real-world or hypothetical phenomenon or system by approximating the behavior of the phenomenon or operation of the system. To more clearly differentiate a simulation from a digital game or visualization, the team made two important distinctions.

First, 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.[1] For the meta-analysis, a computer-based tool was classified as a game if the user needed to complete levels or achievements in order to progress. Typically, a simulation was something that allowed users to be more focused on the behaviors or processes of a specific phenomenon or system than on achieving non-learning-based goals.

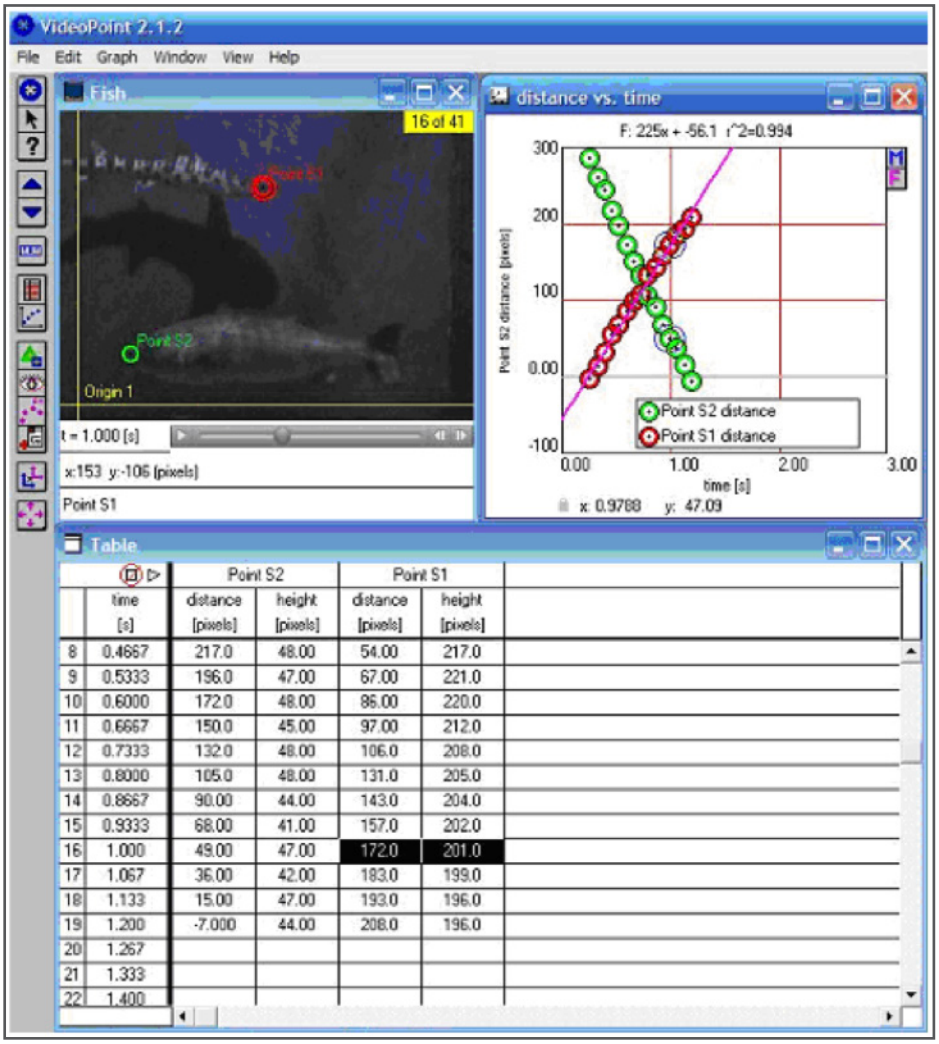
The second distinction was between a simulation and a visualization. This distinction hinges on the important concept of interaction with a scientific model. Simulations, as defined here, must be constructed with an underlying model that is based on some real-world behavior or natural/scientific phenomena (such as models of the ecosystem or simulated animal dissections). The important criterion is that the simulation includes some interactivity on the part of the user, centered usually on inputs and outputs or more generally, the ability to set parameters for modeling the phenomenon or system. 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 this larger GlassLab-Research project. The game/simulation boundary resulted from a discussion between this group and our team to minimize overlap and ensure that neither team overlooked any tools that should be included in our searches. For example, virtual worlds fell along the boundary between simulations and games, and the two groups decided that these should be part of the simulation meta-analysis.

To facilitate our article search and review, we used the term "simulation" as well as other related terms to increase the likelihood that we would identify a range of simulation tools that met our definition. Thus, a research article did not have to explicitly refer to a simulation as a "simulation". The tool described in the article did, however, need to meet our definition. This was especially relevant to the domain of mathematics where often the simulation tools were described as dynamic representations or linked multiple representations.

Exhibit 1 (taken from Ozgun-Koca, 2008) shows an example of a simulation used in a mathematics classroom. In this simulation, students explore the swimming movement of fish by measuring the position of fish over time. The movie, graph, and table are linked to show the fish's positions at one-second intervals. Students can enter information or make changes in any one of the representations (i.e, movie, graph, or table) and see how this changes the others. They can also compare before and after models of fish movement and relate the algebraic form and the graph.

Exhibit 1. An Example of a Simulation in Mathematics from Ozgun-Koca, 2008



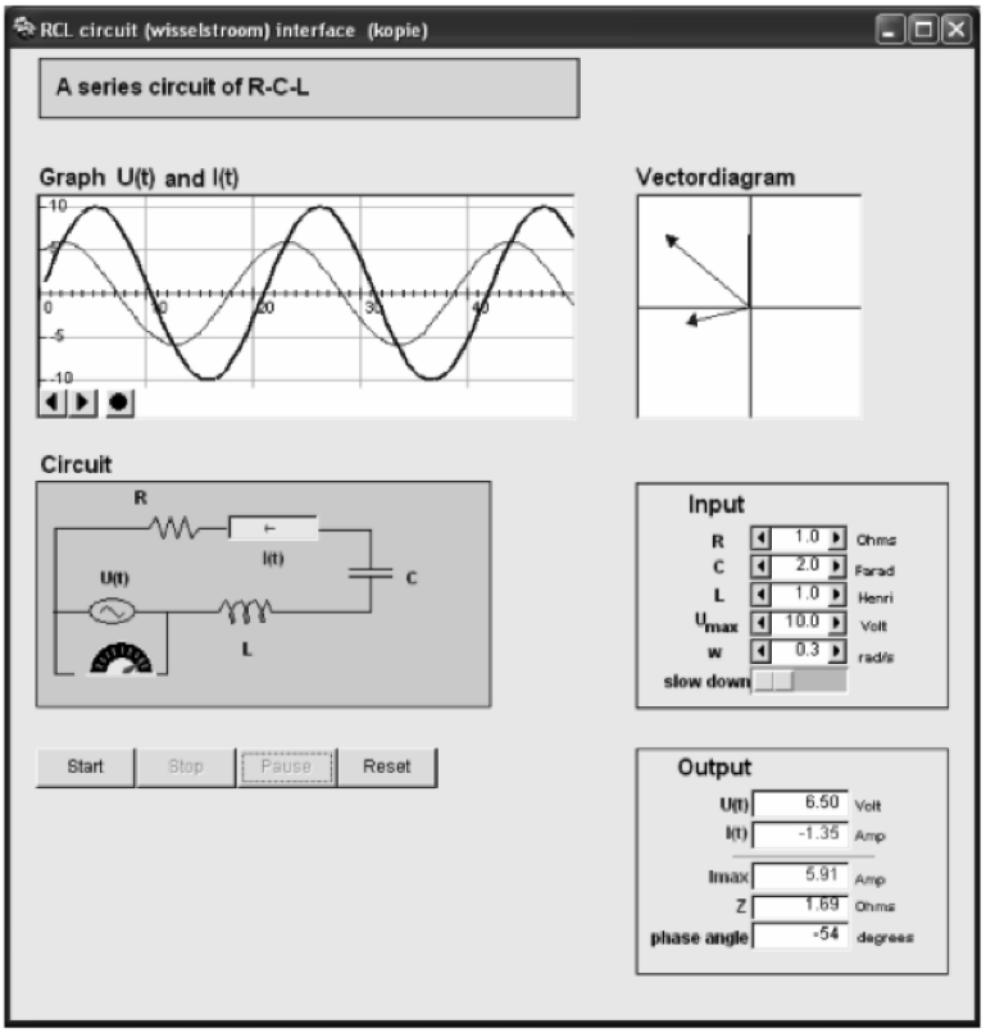
	time [s]	Point S2 distance [pixels]	Point S2 height [pixels]	Point S1 distance [pixels]	Point S1 height [pixels]
8	0.4667	217.0	48.00	54.00	217.0
9	0.5333	196.0	47.00	67.00	221.0
10	0.6000	172.0	48.00	86.00	220.0
11	0.6667	150.0	45.00	97.00	212.0
12	0.7333	132.0	48.00	106.0	208.0
13	0.8000	105.0	48.00	131.0	205.0
14	0.8667	90.00	44.00	143.0	204.0
15	0.9333	68.00	41.00	157.0	202.0
16	1.000	49.00	47.00	172.0	201.0
17	1.067	36.00	42.00	183.0	199.0
18	1.133	15.00	47.00	193.0	196.0
19	1.200	-7.000	44.00	208.0	196.0
20	1.267				
21	1.333				
22	1.400				

Exhibit 2 shows an example from a science simulation in which students investigate electrical circuits (Vreman-de Olde & de Jong, 2006). The interface of the simulation shown here displays a series RCL circuit that students can explore by changing input variables (e.g., resistance and capacitance) of the circuit and relating them to the output variables (e.g., voltage and current). How these variables change the voltage and current are also represented in a graph and a vector diagram.

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.

Exhibit 2. An Example of a Simulation in Science from Vreman-de Olde, & de Jong, 2006



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. For instance, 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.

The study results presented here provide a look at the factors that influence learning science and engineering in computer-based simulations. The final report includes 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 Analysis of Variance (ANOVA). 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. 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 – in this case moving a student from the mean of a normal distribution to one standard deviation above the mean in that distribution.

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 interventions involving simulations in STEM contexts or content in order to align with the GlassLab game developers' objectives. The analysis only included studies with participants in the K–12 grade range (although interventions 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 in the search process in order 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 reference 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 so as to overload our process. 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 actually conducted. Reviewing such a large 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. To try and include mathematics articles that might not use the word "simulation" we included *multiple representation, dynamic representation*, and *linked representation* to the 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.

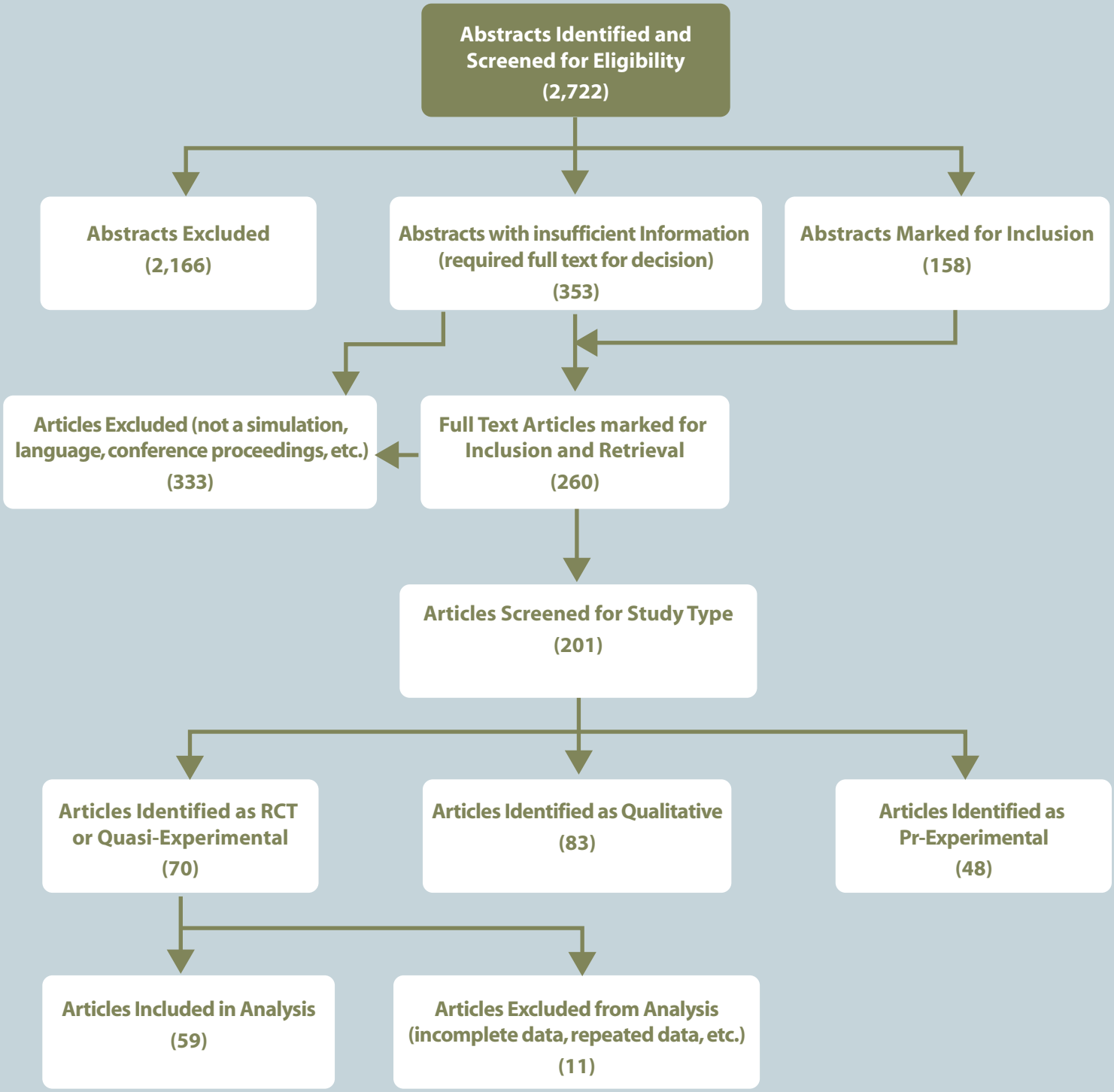
Only articles published between 1991 and 2012 (inclusive) were included in the study. The majority of simulation-based education research studies were conducted during this time, and any studies done before 1991 are likely to concern technologies that are out of date and would not be helpful to contemporary 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 meta-analysis portion, studies needed to include the relevant quantitative information needed for the effect size calculations.

Method Overview

Exhibit 3 presents an overview of the abstract and article retrieval, and coding process. Overall, 2,722 abstracts were reviewed, resulting in full-text retrieval of 260 primary research studies potentially suitable for the analysis. Through a thorough review of full-text documents, 201 studies were retained for further analysis. Of these, 70 were determined to be research articles that included either an experimental (e.g., Randomized Control Trial) or quasi-experimental design. Of those, 11 were determined to contain incomplete or repeated data and were excluded from our analysis. The remaining 59 studies yielded 128 effect sizes, 96 of which were in the achievement outcome category, 17 were in the scientific inquiry and reasoning skills category, and the remaining 15 were in the non-cognitive measures category.

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

Exhibit 3. Abstract and Article Retrieval and Coding Process



Abstract Screening Stage

The abstracts for the 2,722 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 4). 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 review 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 a systematic 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 of 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, 201 (7%) of the original abstracts screened were determined to match all our inclusion criteria and appeared to address one or both of the research questions. For about 350 (13%) of the abstracts screened, 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 80% of the abstracts were excluded for one or more reasons.

Exhibit 4. Abstract Screening Results: Exclusions

Exclusion reason	Number of Abstracts	Percentage of Abstracts
Not K–12 grade range	1007	37.0
Not a research-based article	975	35.8
Simulation is not part of instruction	439	16.1
Does not fit our simulation definition (other)	435	16.0
Review or trend article	141	5.2
Content is not STEM related	126	4.6
Not computer based	150	5.5
Game	66	2.4
Visualization	38	1.4

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

2 http://www.mendeley.com

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

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 read through the full texts 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% ($\kappa = 0.89$). (The list of articles included in the study is in Appendix A.)

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 descriptive purposes and for use in subsequent moderator variable analyses. 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.

Some of the codes were applied at the article or study level (e.g., pertaining to research design or location of the study), whereas others were applied at the effect size level (e.g., 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, platform)
6. Implementation information (setting, curriculum, time/duration/frequency).

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

All of the 59 included articles were coded with the finalized coding scheme by two research team members. A pair of researchers coded each article independently and then met to address any disagreements and determine a final code for each finding.

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.[4] The weights for the fixed-effect model and the random-effects model are different, owing to the theoretical definitions of the models (e.g., Borenstein, Hedges, Higgins & Rothstein, 2010). We used 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 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.

The fixed effect model is then used to compare the categories using the statistic *Q*-Between. *Q*-Between is assessed for significance using the χ^2 sampling distribution with $p - 1$ degrees of freedom. If *Q*-Between is found to be significant (e.g., $\alpha \leq .05$), post hoc analysis may be used to assess differences between categories. A Bonferroni corrected post hoc test is generally used for simple comparisons, where

$$z = \frac{\Delta_{g+}}{SE_{Pooled}}.$$

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

From our review of the literature on computer simulations in K–12 STEM education, we identified three, commonly used outcome measure categories – achievement measures, scientific inquiry and reasoning skills, and non-cognitive measures (including measures such as attitudes). Another key product of our review was the articulation of two guiding research questions for the meta-analysis see the Meta-analysis Results Section of this summary below.

Descriptive Results

The 59 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 and used as descriptive variables to help better understand the pool of articles selected for the study (including demographics of participants, specific simulation topic, etc.). Exhibits 5 (study level) and 6 (effect size level) detail the results of coding for the pertinent descriptive variables.

The 59 studies were found to contain 128 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 three categories of 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 5. Descriptive Results at Study Level (59 Studies)

Variable	Frequency
Location of study	
North America	21
Europe	18
Asia	6
Not indicated	14
STEM domain	
Science	49
Mathematics	6
Engineering	3
Technology	1
Grade level of participants	
Kindergarten– grade 5	8
Grades 6–8	15
Grades 9–12	32
Multiple grade ranges	4

Exhibit 6. Descriptive Results at Effect Size Level (128 Effect Sizes)

Variable	Research Question	
	1: Simulation vs. No Simulation (k = 64)	2: Simulation Plus Enhancement vs. Simulation Alone (k = 64)
Outcome measures		
Achievement	46	50
Scientific Inquiry & Reasoning	6	11
Non-Cognitive	12	3
Simulation type		
Phenomenon simulation	22	41
Virtual lab	19	13
Agent based	5	3
Virtual world	1	1
Other	2	6
Not Indicated	15	0
Assessment delivery mode		
Embedded in simulation	1	15
Tech based but not embedded	1	7
Not tech based	59	27
Not indicated	3	15

Meta-Analysis Results

The Entire Collection

Overall, our systematic search resulted in 2,722 abstracts that were reviewed, resulting in full-text retrieval of 260 primary research studies potentially suitable for the analysis. Through a thorough review and analysis of full-text documents, 59 studies met all of our criteria for inclusion in the met-analysis. These studies yielded 128 effect sizes, 96 of which were in the achievement outcome category, 17 were in the scientific inquiry and reasoning skills category, and the remaining 15 were in the non-cognitive measures category.

Inter-Rater Reliability

At least two trained raters were involved in all stages of the review. Percent agreement and Kappas (when available) pertinent to each stage are:

- Abstract screening—86.1%
- Full-text manuscript inclusion/exclusion—88.35% ($\kappa = 0.77$)
- Effect size comparison decisions (defining experimental and control condition, deciding on the number of effects and which data sources to use)—85.90% ($\kappa = 0.72$)
- Effect size calculation (i.e., accuracy of data extraction and selection and application of equations)—91.44% ($\kappa = 0.83$)

Research Questions

Two research questions were identified in the research literature of educational simulations. Studies either involved the comparison of simulation-based instruction to non-simulation instruction or they involved comparing two different simulation environments. The research questions are defined as follows:

1) What is the difference, in terms of the three outcome types (i.e., content achievement, scientific reasoning and inquiry skills, non-cognitive), between K-12 students who receive simulations as a form of instruction and K-12 students who receive some other kind of instructional treatment? For each included study, the simulation group was designated the experimental condition and the non-simulation group was designated as the control condition.

2) What is the difference, in terms of the three outcome measures, between K-12 students who receive simulations and those who receive the same simulations that are modified with some form of instructional enhancement, such as dynamic representations, meta-cognitive support, or extended feedback? For each included study, the simulations with modifications group was designated the experimental condition and the non-modified simulation group was designated as the control condition.

Interpretation of Results

In the results that follow, there are several aspects of the summary statistics that are important to consider and helpful in interpretation. First, there is the statistical significance ($p < .05$) of the weighted average effect size. This is a formal indication that the average effect size has equaled or exceeded what would be expected by chance (i.e., $g+ > 0$). Second, the magnitude of the effect size, as established by Cohen (1988), can be benchmarked as follows: 1) $0.20 < d < 0.50$ is referred to as a small average effect; 2) $0.50 < d < 0.80$ is referred to as a medium effect) and 3) $d \geq 0.80$ is called a large effect. Valentine and Cooper (2003) warn that these qualitative descriptors may be misleading in fields like Education where smaller effect sizes tend to be the norm.

Another descriptor used here is referred to as U_3 (Cohen, 1988) or the "percentage of scores in the lower-meaned group that are exceeded by the average score in the higher-meaned group" (Valentine & Cooper, 2003, p. 3). For an average effect size of $d = 0.50$, U_3 is approximately 69% of the area under the normal curve. This means that students at the mean of the treatment condition exceeded 69% of students in the control condition. It also means that students at the average of the treatment outperformed students at the average of the control group (i.e., the 50th percentile) by 19% (i.e., 69% – 50% = 19%). We will refer to this difference as an "improvement index" when the treatment condition outperforms the control condition (i.e., + valence for the effect size). For example, for Research Question 1 in this study, an average effect size of $d = 0.50$ would mean that participants in the simulation treatment would have exceeded the performance of students engaged in an equivalent non-simulation activity by 19%. For Research Question 2, this would mean that participants in the treatment condition containing some enhancement would have exceeded average students in the condition that received only simulation by 19%. Again, care needs to be taken in interpreting these percentages because not all distributions of effect sizes are normally distributed, as is presumed in this approach to interpretation.

The variability of a collection of effect sizes is also important. The statistic Q_{Total} is a sum of squares derived from squared deviations around the mean of the distribution that expresses this variability. Q_{Total} is tested for heterogeneity using the Chi-square distribution with $k - 1$ degrees of freedom. I^2 is a statistic, derived from Q_{Total} that indicates the percentage of true heterogeneity in a distribution (i.e., the heterogeneity exceeding chance expectations) and is assessed qualitatively (Higgins, Thompson, Deeks & Altman, 2003) as low ($I^2 < 25\%$), medium ($25\% < I^2 < 75\%$) and high ($I^2 \geq 75\%$). Medium to high levels of heterogeneity indicate that the fixed effect mean is not a good fit to the data and that there possibly are sources of between-study variability (i.e., moderator variables) that can be explored to help describe, more specifically, the nature of the average effect size.

Analyses of Publication Bias, Sensitivity, and Methodological Quality

Outlier and Sensitivity Analysis

Outliers can play a significant role in distorting both the overall mean and variability of a collection of effect sizes. This is especially true for the fixed effect model where the inverse of within study variance is used to give priority to large studies with small standard errors and diminish the effect of smaller studies (Viechtbauer & Cheung, 2010). The random effects model ameliorates this bias somewhat by incorporating average between-study variance (i.e., tau-squared) into the inverse variance weights. However, even with the random effects model, unrealistically large positive or negative effect sizes should be degraded in magnitude or removed from the collection. The following sections detail the outlier analysis for three educational outcomes – achievement, scientific inquiry and reasoning skills and non-cognitive outcomes – across the two research questions.

Achievement Outcomes

The range in the entire collection of achievement outcomes was from g = -0.64 to +4.68. Because of its potentially biasing effect, the large positive value of +4.68 was truncated to the next highest value of +2.81. No adjustments were made to the negative end of the distribution. Also, there were no outliers detected by the one-study removed procedure in Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins & Rothstein, 2011). The average effect size was reduced from $g+$ = 0.57 to $g+$ = 0.55. Thus, the overall random effects weighted average effect size was $g+$ = 0.55 with k = 96 outcomes.

Scientific Inquiry and Reasoning Skills

There were no outliers in this collection, since individual effect sizes ranged from g = -0.82 to +0.88, and based on the one study removed procedure, none exerted a leveraging effect (i.e., extreme effect sizes with large sample sizes) on the weighted average effect size. The overall random effects weighted average effect size was $g+$ = 0.36 with k = 17 outcomes.

Non-Cognitive Outcomes

This collection of outcomes primarily included measures of attitudes, self-efficacy and other similar outcomes found in the literature. There were k = 15 outcomes and no outlying effect sizes. The range was g = -0.82 to 2.40 and the random effects weighted average effect size of $g+$ = 0.66.

Publication Bias Analysis

Analysis of publication bias seeks to determine if a sizable number of studies might have been missed or otherwise not included in a meta-analysis (Rothstein, Sutton, & Borenstein, 2005) and that this number, if found and included, would nullify the average effect. There are various tools for assessing this bias, including the examination of funnel plots (i.e., effect size by standard error) and statistical procedures like classic fail-safe analysis and Orwin's fail-safe procedure. The classic fail-safe procedure is used to determine how many null-effect studies it would take to bring the probability of the average effect to α Orwin's procedure indicates the number of null studies needed to bring the average effect size to some standard of triviality (e.g., $g+$ = 0.10). Duval and Tweedie's (2004) procedure seeks to specify the number of missing effect sizes necessary to achieve symmetry between effect sizes below and or above the mean. It then recalculates $g+$ considering the studies that were imputed.

Achievement Outcomes

Publication bias was analyzed on the complete collection of achievement outcomes across the two questions. For achievement data (k = 96), using the Classic Fail-Safe procedures, 8,909 additional null-effect studies would be necessary to bring the overall probability to α = .05. Additionally, according to Orwin's Fail Safe procedure, 371 null-effect studies would be required to bring g+ to the trivial level of 0.10. According to Duval and Tweedie's Trim and Fill Procedure, no effect sizes should be added to the negative side of the distribution to achieve symmetry under the random effects model. Thus, the random effects weighted average effect size, with k = 96 effects, remains unchanged from $g+$ = 0.55.

Scientific Inquiry and Reasoning Skills

The number of scientific and reasoning skills outcomes was $k = 17$. Using the Classic Fail-Safe procedure, 112 additional null-effect studies would be necessary to bring the overall probability to $\alpha = .05$. Additionally, according to Orwin's Fail Safe procedure, 51 null effect studies would be required to bring $g+$ to the trivial level of 0.10. According to Duval and Tweedie's Trim and Fill Procedure, one imputed value could be added to the negative end of the distribution. Adding this hypothetical effect size would bring symmetry and reduce the random effects model g+ from 0.36 to 0.33. According to this, no significant change would be required in interpreting the random effects weighted average of $g+ = 0.36$.

Non-Cognitive Outcomes

The number of effect sizes representing Non-Cognitive Outcomes is $k = 15$. Using the Classic Fail-Safe procedure, 196 additional null-effect studies would be necessary to bring the overall probability to $\alpha = .05$. Additionally, according to Orwin's Fail Safe procedure, 51 null-effect studies would be required to bring g+ to the trivial level of 0.10. According to Duval and Tweedie's Trim and Fill Procedure, no imputed values were required to achieve symmetry. Thus, the random effects $g+$ for this outcome type remains unchanged at 0.66.

Methodological Quality Assessment

The methodological quality of the research that is included in a meta-analysis is important because of the influence it exerts on the interpretability of the results. While meta-analyses, by definition, do not provide causal explanations, even when all of the studies are of high quality, research design, in particular, affects the clarity of explanations that can be extracted from the study. As a result, we investigated methodological quality and attempted to reduce, as much as possible, its influence as a counter explanation to the research questions under study.

Research Design

A study's research design is arguably its most important characteristic in regards to threats to internal validity. In a randomized control trial (RCT), participants are assigned to groups at random so that selection bias, one of Campbell and Stanley's (1963) most important quality criteria, is reduced or neutralized. In a quasi-experimental design (QED), intact groups (e.g., classes in a school) are used and selection bias is reduced or eliminated through pretesting or matching (Cook & Campbell, 2001). As a first step we removed all pre-experiments that did not include a control condition (i.e., pretest-posttest only designs) or any mechanism for controlling selection bias. Second, in assessing the methodological quality of the

Exhibit 7. Comparison of Randomized Control Trials (RCTs) with Quasi-Experimental Design Studies (QEDs) by Outcome Type

Research Design	*Effect Size and Standard Error*			*Confidence Interval*	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
Achievement Outcomes					
RCTs	32	0.52	0.11	0.29	0.74
QEDs	64	0.57	0.06	0.46	0.68
Q-Between = 0.18, *df* = 1, *p* = .67					
Scientific Inquiry and Reasoning Skills					
RCTs	10	0.37	0.15	0.07	0.67
QEDs	7	0.34	0.15	0.04	0.63
Q-Between = 0.02, *df* =1, *p* = .89)					
Non-Cognitive Outcomes					
RCTs	3	0.21	0.13	-0.05	0.47
QEDs	12	0.80	0.34	0.14	1.46
Q-Between = 2.66, *df* =1, *p* = .10					

*$p < .05$

three collections of outcome measures, levels of the two remaining types of research design (RCTs & QEDs) were compared using the mixed effects moderator variable model. In this approach to moderator variable analysis, *Q*-Between (i.e., the *Q*-value indicating the difference between groups) is tested for significance.

Exhibit 7 shows the overall weighted average effect size for three collections of effect sizes, separated by research design, randomized control trials (RCTs) and quasi-experimental designs (QED).

For all three outcome types, the test of differences among categories (*Q*-Between) indicates that RCTs and QEDs were not significantly different from one another. Effectively, this confirms that the type of research design will not influence the interpretation of the overall weighted average effects.

Measurement Validity and Reliability

According to Valentine and Cooper (2008) the quality of measurement in primary studies is an important consideration to be taken into account by a meta-analyst. Two aspects of measurement were coded, instrument validity and instrument reliability, and then subsequently combined into a single coded variable representing the joint technical quality of the instrumentation in simulation studies. Due to the small number of effect sizes and the limited number of categories for scientific inquiry and reasoning skills and non-cognitive outcomes, only achievement outcomes were subjected to scrutiny. (For more information on the breakdown of types of validity and reliability evidence see the Assessment section.)

Two forms of validity were identified from the literature – expert panel review and evaluation and correlational corroboration. Four coded levels of validity resulted: 1) expert panel review and evaluation; 2) correlational corroboration; 3) both; and 4) not indicated (NI). The "both" category was considered stronger evidence of instrument validity. In this case NI was left in the comparison to determine whether validity results that were reported differed from those that were not reported. *Q*-Between revealed no significant differences among these categories (*Q*-Between = 4.99, df = 3, p = .17). Similarly, instrument reliability was tested by using three categories: 1) the reliability coefficient (such as Cronbach's alpha or Cohen's kappa) was relatively high (above .75); 2) the reliability coefficient was between .4 and .75; and 3) the reliability coefficient was not indicated. Again, *Q*-Between revealed no significant differences among these categories (*Q*-Between = 1.25, df = 2, p = .54) as shown in Exhibit 8.

Other Measures of Methodological Quality

There were five other coded study features that can have an impact on methodological study quality 1) source of the outcome measure; 2) effect size extraction procedure (i.e., the quality of the statistical information available in studies); 3) instructor equivalence/non-equivalence; 4) learning materials equivalence/non-equivalence; and 5) time-on-task equivalence/non-equivalence. Each of these was considered to be an important indicator of methodological quality, but since none was considered as important as Type of Research Design (see Exhibit 7),

Exhibit 8. Comparison of Levels of the Combined Indicators of Test Validity and Test Reliability for Achievement Outcomes

Categories	*Effect Size and Standard Error*			*Confidence Interval*	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
Reliability and Validity Indicators	12	0.72	0.12	0.47	0.96
Reliability or Validity Indicators	37	0.61	0.08	0.45	0.78
Not Indicated (NI)	47	0.47	0.08	0.31	0.62
Q-Between = 2.66, df = 1, p = .10					

*$p < .05$

these five elements were combined into an ordinal level composite index of methodological quality (Abrami & Bernard, 2013). Each variable had two levels, one ("1") designating higher quality (i.e., the absence of this threat to internal validity) and one ("0") meaning lower level (i.e., possible presence of the threat to internal validity). For instance, in a particular study if two teachers each taught different groups (e.g., the treatment and control conditions), individual teacher differences might produce an alternative explanation of results. This methodological study feature would be coded "0," on the corresponding categories indicating this possible biasing effect. The full scale of five variables, then, ranged from a low of 0 (i.e., minimum quality) to a high of 5 (i.e., maximum quality). No studies ended up with a value of 0, 9 studies had a value of 1, 27 studies had a value of 2, 54 had a 3, 20 studies had a 4, and 18 studies had a value of 5. The average value for these studies was 3.09.

The results of random effects (method of moments) meta-regression, run with inverse variance weighted effect sizes as the outcome measure and the methodological quality index as the predictor, yielded the results in Exhibit 9. No significant prediction was observed in any of the three outcome types. Only one outcome type contained any suggestion of bias due to the combination of these methodological features. In the outcome category *Scientific Inquiry and Reasoning Skills* there is a positive slope, suggesting that higher quality studies are higher on the effects in this outcome type, but it not significant (Q-Regression = 3.11, df = 1, p = .08). All three meta-regression models were homogenous, based on the significance level of Q-Residual ($p > .25$).

Overall, these analyses of the methodological quality of the three sub-collections of outcomes reveal no weaknesses that would diminish the straight forward interpretations of the results. In the next section, we proceed to the analysis by outcome measure and by research question.

Exhibit 9. Meta-regression of Effect Sizes with Methodological Quality Index

Research Design	*Slope, Intercept and Standard Error*			*Confidence Interval*	
	k	b_Y	*SE*	*Lower 95th*	*Upper 95th*
Achievement Outcomes					
Slope	96	0.05	0.05	-0.05	0.14
Intercept		0.41	0.16	0.10	0.72
Q-Regression = 0.90, df = 1, p = .34 Q-Residual = 99.40, df = 94, p = .33					
Scientific Inquiry and Reasoning Skills					
Slope	17	0.18	0.10	-0.02	0.39
Intercept		-0.24	0.36	-0.95	0.46
Q-Regression = 3.11, df = 1, p = .0 Q-Residual = 15.36, df = 15, p = .43					
Non-Cognitive Outcomes					
Slope	15	-0.41	0.30	-0.99	0.17
Intercept		1.80	0.85	0.12	3.47
Q-Regression = 1.95, df = 1, p = .16 Q-Residual = 14.51, df = 13, p = .34					

*$p < .05$

Analysis by Outcome Type and Research Question

Research Questions

As stated above, two research questions were identified in order to classify the studies. These research questions are:

1) What is the difference, in terms of the three 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? The simulation condition was designated the experimental condition and the other was designated as the control condition.

2) What is the difference, in terms of the three outcome measures, between K-12 students who receive simulations and those who receive the same simulations that are modified with some form of instructional enhancement, such as dynamic representations, meta-cognitive support, or extended feedback? For each included study, the simulations with modifications group was designated the experimental condition and the non-modified simulation group was designated as the control condition.

In addition, three outcome variables were identified: achievement outcomes, scientific inquiry and reasoning skills, and non-cognitive outcomes. Exhibits 10, 11 and 12 show the results of the analyses by outcome type and research question. For achievement outcomes (Exhibit 10), both questions produced moderately high (i.e., by Cohen's criteria) weighted average effect sizes (random effects). The improvement indices of these effect sizes were 23% and 19%, respectively. The weighted average effect sizes of both collections were significantly greater than zero and both were significantly heterogeneous under the fixed effect model. The heterogeneity analyses,

Exhibit 10. Summary Statistics for Question 1 and Question 2 for Achievement Outcomes

Research Questions 1 and 2	Effect Size and Standard Error			Confidence Interval	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
Question 1 ($N = 2{,}947$): Simulation vs. No Simulation	46	0.62*	0.09	0.45	0.79
Heterogeneity	$Q\text{-}Total = 209.80, df = 45, p < .001, I^2 = 78.55$				
Question 2: ($N = 3{,}342$) Simulation Plus Modification vs. Non-Modified Simulation	50	0.49*	0.06	0.36	0.61
Heterogeneity	$Q\text{-}Total = 131.24, df = 49, p < .001, I^2 = 62.66$				

*$p < .05$

Exhibit 11. Summary Statistics for Question 1 and Question 2 for Scientific Inquiry and Reasoning Skills

Research Questions 1 and 2	Effect Size and Standard Error			Confidence Interval	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
Question 1 ($N = 347$): Simulation vs. No Simulation	6	0.26	0.15	-0.03	0.55
Heterogeneity	$Q\text{-}Total = 7.85, df = 5, p = .17, I^2 = 36.32$				
Question 2: ($N = 689$) Simulation Plus Modification vs. Non-Modified Simulation	11	0.41*	0.15	0.13	0.70
Heterogeneity	$Q\text{-}Total = 32.03, df = 10, p < .001, I^2 = 68.77$				

*$p < .05$

based on the fixed effect model, were significant and the I^2s were moderate to high. Both are candidates for additional moderator variable analysis.

Results for the outcomes referred to as Scientific Inquiry and Reasoning Skills produced a somewhat more modest effect for Research Question 1 based on $k = 6$ effect sizes, resulting in a weighted average effect size (random effects) of $g+ = 0.26$ that was low (Improvement Index = 10%) and not significantly greater than zero. The small distribution was statistically homogeneous, but this is not an unusual finding for small collections, since there is a known bias in *Q*-Total in such circumstances (Higgins et al. 2003). Predictably, I^2 is also low since it is based on *Q*-Total.

Research Question 2, however, based on 11 effect sizes fared better, producing a weighted average effect size (random effects) of $g+ = 0.41$. It was significantly greater than chance expectations (Improvement Index = 16%), and the collection was significantly heterogeneous. The small number of studies made it unlikely that this outcome measure would yield any important findings in moderator variable analysis.

The results for *Non-Cognitive Outcomes* (Exhibit 12) were similar to the achievement results but based on many fewer effect sizes. Question 1 produced the highest of all weighted average effect size (Improvement Index = 25.5%) and it also contained the highest level of between-study heterogeneity (i.e., indicating that the studies differed greatly). With $k = 12$ effect sizes and such wide variability, it is unlikely that this outcome type would yield informative results in moderator variable analysis.

Question 2 for *Non-Cognitive Outcomes* produced only $k = 3$ studies, too few to provide a reasonable assessment for this question and outcome type. Interestingly, unlike Question 1 for *Non-Cognitive Outcomes*, these results were nearly homogeneous. Again, this is most likely to be the result of the very small number of effect sizes.

Exhibit 12. Summary Statistics for Question 1 and Question 2 for Non-Cognitive Outcomes

Research Questions 1 and 2	*Effect Size and Standard Error*			*Confidence Interval*	
	k	***g+***	***SE***	***Lower 95th***	***Upper 95th***
Question 1 ($N = 663$): Simulation vs. No Simulation	12	0.69*	0.33	0.05	1.33
Heterogeneity	*Q-Total* = 155.45, *df* = 11, $p < .001$, $I^2 = 92.92$				
Question 2: ($N = 205$) Simulation Plus Modification vs. Non-Modified Simulation	3	0.52*	0.25	0.03	1.02
Heterogeneity	*Q-Total* = 5.74, *df* = 2, $p = .06$, $I^2 = 65.13$				

*$p < .05$

Moderator Variable Analysis – Demographic and Substantive

Moderator variable analysis is conducted in an attempt to uncover underlying patterns of difference among coded levels of moderator variables (i.e., study characteristics: demographic and substantive) that might help to explain overall between-study variability in average effect size. When the random effects model is used for the primary analysis, it is appropriate to use the mixed effects model for follow-up analysis. The mixed effects model, as the name implies, uses both fixed and random analytical approaches to come to a conclusion about the difference between coded categories of a moderator variable. The random effects model is used to produce weighted average effect sizes within categories of the moderator. The weighting approach is identical to that performed on the entire collection, using the same random weights that contain the two types of variance (i.e., within group variance for each study – v_i – and average between group variance $-\tau^2$ – for each subgroup). As discussed in the Methods section, *Q*-Between is assessed for significance. If *Q*-Between is found to be significant (e.g., $\alpha \leq .05$), post hoc analysis may be used to assess differences between categories.

Research Question 1 (Simulation versus No Simulation)

Achievement Outcomes

Exhibit 13 shows the breakdown of studies between math and science. The weighted average effect size for science studies is significantly greater than zero, but math is not. Math and science are significantly different from one another, but the math category contains only four cases and the *g*+ is negative. Based on the difference in studies plus the large difference in average effects and their potentially biasing effects the decision was made to separate math and science in further analyses related to Research Question 1 for achievement outcomes. In the remainder of this section, only science simulations will be examined.

Demographic Moderator Variables for Research Question 1

Exhibit 14 contains the results of the demographic study features moderator variables with math studies removed. Categories of grade level are not significantly different from each other, although all three weighted average effect sizes are significantly different from zero.

Exhibit 13. Comparison of Levels of Math and Science for Research Question 1

Research Questions 1 and 2	*Effect Size and Standard Error*			*Confidence Interval*	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
STEM Subject					
Math	4	-0.15	0.19	-0.52	0.22
Science	42	0.67*	0.09	0.50	0.84
Q-Total = 15.28, *df* = 1, *p* < .001					

*$p < .05$

These results suggest that the usefulness of simulations is essentially uniform, but it is noted that there are only three effect sizes for the Grade K-5 grade category. This small sample size contributes greatly to the fact that, in spite of its large effect size $g+ = 1.42$, this category suffers from an underpowered comparison. Among the remaining demographic variables, there is none in which levels are discriminated from one another. By and large, the effect sizes for the levels in each of these categories are large enough to be significantly different from zero.

The effect sizes coded NI (i.e., not indicated) in all of the following analyses are shown, with all statistics included. However, NI is not included in the test of difference (*Q*-Between) between levels of moderator variables, as it cannot be meaningfully interpreted.

Substantive moderator variables for Question 1

The next set of analyses, summarized in Exhibit 15, is coded categories of substantive moderator variable. Substantive in this context relates to the fact that these variables are instructional characteristics that to some extent can be manipulated by the teacher. While the average effect size in most levels within each variable category is significant, only one variable (Number of Sessions) significantly discriminates among levels. Post hoc contrasts (Bonferroni's method, Hedges & Olkin, 1985) revealed that the category "four – six sessions" produced a significantly higher weighted average effect size ($z = 2.53$, $p < .05$) than the other three categories ("one session," "two - three sessions" and "more than eleven sessions").

Exhibit 14. Comparison of Levels of Demographic Moderator Variables for Question 1 for Achievement Outcomes With Math Studies Removed

Levels	Slope, Intercept and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Grade Ranges (Science only)					
Grades K-5	3	1.42*	0.72	0.003	2.84
Grades 6-8	12	0.64*	0.19	0.27	1.02
Grades 9-12	26	0.63*	0.10	0.44	0.82
Q-Between = 1.19, *df* = 2, *p* = .55					
Note: One study with multiple grade ranges was excluded here.					
Language of Instruction (Science only)					
English	20	0.78*	0.11	0.56	2.99
Non-English	11	0.40	0.21	-0.01	0.81
NI	11	0.74*	0.14	0.46	1.02
Q-Between = 2.10, *df* = 1, *p* = .11 (With category NI excluded)					
Regions of the World (Science only)					
U.S.	15	0.79*	0.13	0.53	1.04
Europe	7	0.51	0.29	-0.06	1.08
Asia	5	0.84*	0.28	0.30	1.38
Middle East	11	0.48*	0.09	0.31	0.64
NI	4	0.88*	0.13	0.63	1.13
Intercept		1.80	0.85	0.12	3.47
Q-Between = 4.85, *df* = 3, *p* = .18 (With category NI excluded)					

*$p < .01$

Exhibit 15. Comparison of Levels of Substantive Moderator Variables for Question 1 for Achievement Outcomes

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Simulation Type (Science only) (Removed Other = 2; NI = 5; Virtual World = 1)					
Agent-based	5	0.78*	0.11	0.56	1.00
Phenomenon-based	17	0.71*	0.15	0.42	1.01
Virtual Labs	12	0.68*	0.17	0.34	1.01
Q-Between = 0.31, *df* = 2, *p* = .86					
Collaborative Settings (Science only) (Treatment only)					
Collaboration Required	7	0.95*	0.19	0.57	1.33
Collaboration Optional	7	0.59*	0.27	0.07	1.11
No Collaboration	7	0.70*	0.25	0.22	1.19
NI	21	0.58*	0.11	0.37	0.80
Q-Between = 1.33, *df* = 2, *p* = .52 (With category NI excluded)					
Group Work (Science only) (Treatment only)					
Individual Work	13	0.50*	0.16	0.20	0.81
Dyads	8	0.80*	0.23	0.36	1.25
Small Groups	10	0.88*	0.20	0.49	1.28
NI	11	0.60*	0.08	0.45	0.75
Q-Between = 2.58, *df* = 2, *p* = .28 (With category NI excluded)					
Group Work (Science only) (Treatment only)					
Individual Work	13	0.50*	0.16	0.20	0.81
Group Work (Dyads + Small Groups)	18	0.85*	0.15	0.55	1.14
Q-Between = 2.46, *df* = 1, *p* = .12					
Flexibility (Science only) (Treatment only)					
Free Form	5	0.58*	0.17	0.25	0.90
Some Structure	15	0.58*	0.17	0.25	0.91
Very Structured	9	0.89*	0.19	0.52	1.26
NI	13	0.68*	0.14	0.40	0.80
Q-Between = 1.97, *df* = 2, *p* = .37 (With category NI excluded)					
Curriculum (Science only) (Treatment only)					
Simulation Embedded	22	0.81*	0.12	0.58	1.04
Stand Alone/related	12	0.53*	0.18	0.18	0.87
Stand Alone/not related	2	0.20	0.32	-0.43	0.82
NI	6	0.56*	0.13	0.31	0.81
Q-Between = 4.33, *df* = 2, *p* = .12 (With category NI excluded)					

*$p < .05$

Exhibit 15. Comparison of Levels of Substantive Moderator Variables for Question 1 for Achievement Outcomes (Continued)

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Number of Sessions (Science only)					
One Session	11	0.49*	0.16	0.17	0.81
Two to Three	3	0.39*	0.13	0.14	0.64
Four - Six	9	0.91*	0.12	0.68	1.13
Eleven +	8	0.44*	0.10	0.23	0.64
NI	11	0.94*	0.27	0.41	1.47
Q-Between = 12.32, *df* = 3, *p* = .006 (Treatment Only with NI excluded)					
Session Duration (without Math)					
15-30 min.	3	0.63*	0.21	0.21	1.04
40-50 min.	16	0.72*	0.12	0.49	0.95
60-80 min	2	0.46	0.27	-0.07	0.99
90+ min.	5	0.60	0.35	-0.08	1.27
NI	16	0.69*	0.16	0.37	1.00
Q-Between = 0.87, *df* = 3, *p* = .83 (With category NI excluded)					
Total Duration (Constructed: # sessions x session time) (without Math)					
Under 60 min.	8	0.70*	0.12	0.45	0.94
One to five hrs.	8	0.50*	0.16	0.19	0.82
Over 5 hrs.	7	0.67*	0.22	0.24	1.10
NI	19	0.74*	0.16	0.43	1.05
Q-Between = 0.96, *df* = 2, *p* = .62 (With category NI excluded)					

**p < .05*

Non-Cognitive Outcomes

Demographic Moderator Variables for Question 1

Exhibit 16. Comparison of Levels of Demographic Moderator Variables for Question 1 for Non-Cognitive Outcomes

Levels	Slope, Intercept and Standard Error			Confidence Interval	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
Grade Ranges					
Grades 6 – 8	2	-0.40*	0.33	-1.05	0.26
Grades 9 – 12	10	0.92*	0.35	0.24	1.60
Q-Between = 7.46, *df* = 1, *p* = .006					
Language of Instruction					
English	4	-0.21	0.14	-0.49	0.07
Non-English	7	1.48*	0.29	0.90	2.06
NI	1	-0.71	0.19	-1.07	-0.34
Q-Between = 26.96, *df* = 1, *p* < .001 (With category NI excluded)					
Regions of the World (Science only.)					
U.S.	4	-0.21	0.14	-0.49	0.07
Europe	2	1.20	1.19	-1.14	3.53
Turkey	5	1.59*	0.25	1.11	2.07
NI	1	-0.71	0.19	-1.07	-0.34
Q-Between = 41.08, *df* = 2, *p* <.001 (With category NI excluded)					

**p < .05*

Substantive Moderator Variables for Question 1

Exhibit 17. Comparison of Levels of Substantive Moderator Variables for Question 1 for Non-Cognitive Outcomes

Levels	*Effect Size and Standard Error*			*Confidence Interval*	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Simulation Type					
Phenomenon-based	1	-0.71	0.19	-1.07	-0.34
Virtual Labs	6	0.15	0.33	-0.50	0.79
NI	5	1.59	0.25	1.11	2.07
Q-Between N/A					
Assessment Delivery Mode					
Not Technology-based	11	0.51*	0.11	0.30	0.71
Technology-based Embedded	1	0.67*	0.16	0.37	1.02
Q-Between N/A					
Collaborative Settings (Treatment only; [1]Required *k* = 1)					
[1]Collaboration Required/Optional	4	0.57	0.60	-0.60	1.74
No Collaboration	2	-0.13	0.16	-0.45	0.18
NI	6	1.09	0.58	-0.04	2.21
Q-Between = 5.18, *df* = 2, *p* = .08 (Treatment only with NI included) *Q*-Between = 1.31, *df* = 1, *p* = .25 (Treatment only with NI excluded)					
Group Work (Treatment only)					
Individual Work	4	0.47	0.45	-0.41	1.34
Dyads	2	1.58*	0.50	0.61	2.55
Small Groups	2	-0.48	0.30	-1.06	0.11
NI	4	1.04*	0.75	-0.43	2.51
Q-Between = 14.59, *df* = 3,*p* = .002 (Treatment only with NI included) *Q*-Between = 13.34 , *df* = 2, *p* = .001 (Treatment only with NI excluded)					
Flexibility (Treatment only)					
Free Form	1	-.040	0.26	-0.55	0.48
T. Some Structure	3	0.87*	0.65	-0.40	2.13
T. Very Structured	2	-0.22*	0.18	-0.56	0.13
NI	6	1.09*	0.58	-0.04	2.21
Q-Between = 2.61, *df* = 1, *p* = .11 (Treatment only with Free Form & NI excluded)					
Curriculum (Treatment only)					
Simulation Embedded	3	0.99*	0.67	-0.32	2.29
Stand Alone/related	4	0.32*	0.64	-0.94	1.57
Stand Alone/not related	1	-0.04*	0.26	-0.55	0.48
NI	4	1.04*	0.75	-0.43	2.51
Q-Between = .53, *df* = 1, p = .47 (Treatment only with "not related" & NI excluded)					

*$p < .05$

Exhibit 17. Comparison of Levels of Substantive Moderator Variables for Question 1 for Non-Cognitive Outcomes (Continued)

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Number of Sessions					
One Session	4	-0.21	0.14	-0.49	0.07
Four-Six	1	-0.71*	0.19	-1.07	-0.34
Eleven +	5	1.59*	0.25	1.11	2.07
NI	2	1.20	1.19	-1.14	3.53
Q-Between = 40.44, *df* = 1, *p* < .001 (With category "Four - Six" & NI excluded					
Session Duration (In minutes)					
15-30 min.	2	-0.13	0.16	-0.45	0.18
40-50 min.	5	0.30	0.61	-0.90	1.49
60-80 min	2	1.20	1.19	-1.14	3.53
NI	3	1.53*	0.25	1.05	2.01
Q-Between = 1.64, *df* = 2, *p* = .44 (With category NI excluded)					
Total Duration (Constructed: # sessions x session time)					
Under 60 min.	4	-0.21	0.14	-0.49	0.07
One to five hrs.	1	-0.71*	0.19	-1.07	-0.34
Over 5 hrs.	2	1.58*	0.50	0.61	2.55
NI	5	1.45*	0.43	0.61	2.28
Q-Between = 12.15, *df* = 1, *p* < .001 (With category "One to five hrs." & NI excluded)					

*$p < .05$

Research Question 2 (Modified simulation versus non-modified simulation)

Achievement Outcomes: Science and Math

Research Question 2 involves a comparison between a non-modified simulation (the control condition) and a simulation with some instructional modification (the treatment condition). The results of the analyses of demographic moderator variables are shown in Exhibit 19 below. As before, the comparison of math and science comes first. From the table, it is clear that a difference exists between the analysis of this question and the analysis of Question 1. For research question 2 while there is still a significant difference between the effect sizes for math and science, the average effect size is positive. Also, there are more math studies for this research question than there were for the previous research question and so this report contains two analyses for Question 2. The first analysis is the two subject areas (i.e., math and science) combined for demographic and substantive moderators (Exhibit 19 and Exhibit 20). The second analysis (Exhibit 21 and Exhibit 22) deals with science data only (as it was done previously for Research Question 1).

Exhibit 18. Comparison of Levels of Math and Science for Research Question 2 for Achievement Outcomes

Research Questions 1 and 2	Effect Size and Standard Error			Confidence Interval	
	k	*g+*	*SE*	*Lower 95th*	*Upper 95th*
STEM Subject					
Math	10	0.26	0.17	-0.08	0.59
Science	40	0.53*	0.07	0.39	0.66
Q-Total = 2.15 *df* = 1, *p* < .001					

*$p < .05$

Demographic Moderator Variables for Research Question 2

Exhibit 19. Comparison of Levels of Demographic Moderator Variables for Research Question 2 for Achievement Outcomes (Math and Science Combined)

Levels	Slope, Intercept and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Grade Ranges					
Grades K-5	6	0.32	0.24	-0.14	0.78
Grades 6-8	11	0.51*	0.11	0.29	0.72
Grades 9-12	25	0.51*	0.10	0.31	0.71
Multiple	8	0.51*	0.19	0.15	0.88
Q-Between = 0.59, *df* = 3, *p* = .99					
Language of Instruction					
English	20	0.45*	0.11	0.25	0.66
Non-English	11	0.41*	0.12	0.18	0.64
NI	19	0.60*	0.12	0.37	0.83
Q-Between = 0.07, *df* = 1, *p* = .79 (With category NI excluded)					
Regions of the World					
U.S.	15	0.45*	0.12	0.21	0.69
Europe	9	0.55*	0.15	0.25	0.84
Asia & Turkey	8	0.61*	0.18	0.26	0.95
NI	18	0.43*	0.10	0.23	0.63
Q-Between = 0.62, *df* = 2, *p* = .73 (With category NI excluded)					

*$p < .05$

Substantive Moderator Variables for Research Question 2

Exhibit 20 below summarizes the findings for substantive moderator variables. As noted below, for most of the following moderator variable analyses, the coding of the treatment condition only was used. In the cases where the variable might be different between the two conditions for a particular study, this information was almost always contained in the Nature of Modification variable. The moderator variable Assessment Delivery Mode is the only substantive moderator variable that differentiates between levels.

Exhibit 20. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Achievement Outcomes (Math and Science Combined)

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Simulation Type[1]					
Agent-based	3	0.94*	0.37	0.21	1.67
Phenomenon-based	29	0.41*	0.09	0.24	0.58
Virtual Labs	11	0.58*	0.13	0.34	0.83
Q-Between = 2.88, *df* = 2, *p* = .24					
Notes: [1]Removed: Other *k* = 3; NI *k* = 3; Virtual World *k* = 1.					
Nature of Modification to Treatment					
Representations	19	0.32*	0.11	0.10	0.53
Scaffolding	19	0.60*	0.11	0.38	0.82
Cooperative Learning	4	0.69*	0.16	0.37	1.02
Additional Real Lab	2	0.39	0.32	-0.25	1.02
Haptic	3	0.43*	0.12	0.19	0.66
Feedback	2	0.54*	0.27	0.01	1.08
Q-Between = 5.38, *df* = 5, *p* = .37 (All above variables in, except Cultural, *k* = 1)					
Assessment Delivery Mode					
Not Technology-based	27	0.41*	0.09	0.23	0.59
Technology-based Embedded	5	0.67*	0.15	0.38	0.97
Technology-based Not Embedded	4	0.11	0.09	-0.07	0.28
NI	14	0.74	0.13	0.49	0.99
Q-Between = 12.42, *df* = 2, *p* = .002 (With category NI excluded)					
Collaborative Settings[2] (Treatment only)					
Collaboration Required/Optional	19	0.32*	0.10	0.12	0.52
No Collaboration	13	0.47*	0.13	0.23	0.72
NI	18	0.68*	0.11	0.47	0.89
Q-Between = 0.91, *df* = 1, *p* = .34 (Treatment only with NI excluded)					
Notes: [2]Removed: Required *k* = 1					
Group Work (Treatment only)					
Individual Work	20	0.45*	0.10	0.26	0.64
Dyads	8	0.30	0.22	-0.14	0.74
Small Groups	9	0.42*	0.12	0.18	0.66
NI	13	0.72*	0.11	0.50	0.94
Q-Between = .38, *df* = 1, p = .11 (Treatment only with NI excluded)					

*$p < .05$

Exhibit 20. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Achievement Outcomes (Math and Science Combined) (Continued)

Levels	*Effect Size and Standard Error*			*Confidence Interval*	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Flexibility (Treatment only)					
Some Structure	24	0.45*	0.08	0.29	0.61
Very Structured	11	0.74*	0.17	0.41	1.06
NI	15	0.40*	0.12	0.15	0.64
Q-Between = 2.49, *df* = 1, *p* = .11 (Treatment only with NI excluded)					
Curriculum (Treatment only)					
Simulation Embedded	11	0.65*	0.08	0.49	0.80
Stand Alone/related	25	0.44*	0.11	0.23	0.65
Stand Alone/not related	5	0.33*	0.15	0.05	0.62
NI	9	0.47*	0.16	0.15	0.79
Q-Between = 4.58, *df* = 2, *p* = .10 (Treatment only with NI excluded)					
Number of Sessions					
One Session	16	0.45*	0.10	0.26	0.64
Two to Three	12	0.28	0.16	-0.03	0.59
Four - Six	10	0.76*	0.13	0.51	1.01
Eleven +	3	0.87*	0.33	0.22	1.52
NI	9	0.39*	0.12	0.15	0.62
Q-Between = 7.41, *df* = 3, *p* = .06 (With category NI excluded)					
Session Duration (In minutes)					
15-30 min.	4	0.22	0.17	-0.11	0.54
40-50 min.	14	0.62*	0.14	0.34	0.90
60-80 min	10	0.62*	0.11	0.41	0.83
90+ min.	3	0.54*	0.20	0.15	0.93
NI	19	0.39*	0.11	0.17	0.60
Q-Between = 4.66, *df* = 3, *p* = .20 (With category NI excluded)					
Total Duration (Constructed: # sessions x session time)					
Under 60 min.	7	0.35*	0.15	0.05	0.65
One to five hrs.	19	0.59*	0.11	0.37	0.80
Over 5 hrs.	5	0.68*	0.14	0.41	0.95
NI	19	0.39*	0.11	0.17	0.60
Q-Between = 2.74, *df* = 2, p = .25 (With category NI excluded)					

*$p < .05$

Exhibit 21. Comparison of Levels of Demographic Moderator Variables for Research Question 2 for Achievement Outcomes

Levels	Slope, Intercept and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Grade Ranges (Science Only)					
Grades K-5	5	0.43	0.25	-0.07	0.92
Grades 6-8	11	0.51*	0.11	0.29	0.72
Grades 9-12	17	0.58*	0.12	0.35	0.81
Multiple	7	0.58*	0.21	0.17	0.99
Q-Between = 0.44, *df* = 3, *p* = .93					
Language of Instruction (Science Only)					
English	17	0.46*	0.11	0.24	0.67
Non-English	11	0.41*	0.12	0.18	0.64
NI	12	0.80*	0.11	0.57	1.02
Q-Between = 0.08, *df* = 1, *p* = .77 (With category NI excluded)					
Regions of the World (Science Only)					
U.S.	14	0.40*	0.12	0.16	0.63
Europe	9	0.55*	0.15	0.25	0.84
Asia & Turkey	8	0.61*	0.18	0.26	0.95
NI	9	0.64*	0.10	0.44	0.84
Q-Between = 1.17, *df* = 2, *p* = .56 (With category NI excluded)					

*$p < .05$

Achievement Outcomes: Science Only

This section reports the results of moderator variable analysis for science (and engineering) data only, thus mirroring the approach taken with the Research Question 1 analysis. There are 38 science effect sizes and 2 engineering effect sizes in this set of analyses.

Demographic Moderator Variables for Research Question 2

Exhibit 22. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Achievement Outcomes

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Simulation Type[1] (Science only.)					
Agent-based	3	0.94*	0.37	0.21	1.67
Phenomenon-based	29	0.41*	0.09	0.24	0.58
Virtual Labs	11	0.58*	0.13	0.34	0.83
Q-Between = 1.88, *df* = 2, *p* = .39					
Notes: [1]Removed: Other *k* = 3; Virtual World *k* = 1.					
Nature of Modification to Treatment					
Representations	10	0.43*	0.14	0.16	0.69
Scaffolding	20	0.60*	0.11	0.38	0.82
Cooperative Learning	4	0.69*	0.16	0.37	1.02
Additional Real Lab	2	0.39	0.32	-0.25	1.02
Haptic	3	0.43*	0.12	0.19	0.66
Feedback	1	0.54*	0.27	0.01	1.08
Q-Between = 2.74, *df* = 4, *p* = .60					
Assessment Delivery Mode					
Not Technology-based	19	0.51*	0.11	0.30	0.71
Technology-based Embedded	4	0.67*	0.16	0.37	1.02
Technology-based Not Embedded	4	0.11	0.09	-0.07	0.28
NI	13	0.70	0.13	0.45	0.96
Q-Between = 13.96, *df* = 2, *p* = .001 (With category NI excluded)					
Collaborative Settings (Treatment only; [1]Required *k* = 1)					
[1]Collaboration Required/Optional	13	0.39*	0.12	0.15	0.62
No Collaboration	10	0.47*	0.13	0.22	0.72
NI	17	0.68*	0.11	0.46	0.90
Q-Between = 0.22, *df* = 1, *p* = .64 (Treatment only with NI excluded)					
Group Work (Treatment only)					
Individual Work	16	0.49*	0.11	0.28	0.71
Dyads	4	0.56	0.41	-0.24	1.35
Small Groups	8	0.43*	0.13	0.17	0.68
NI	12	0.68*	0.11	0.46	0.90
Q-Between = .22, *df* = 2, p = .90 (Treatment only with NI excluded)					

*$p < .05$

Exhibit 22. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Achievement Outcomes (Continued)

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Flexibility (Treatment only)					
Some Structure	22	0.42*	0.08	0.26	0.58
Very Structured	9	0.88*	0.15	0.58	1.17
NI	9	0.50*	0.16	0.19	0.81
Q-Between = 7.00, *df* = 1, *p* = .008 (Treatment only with NI excluded)					
Curriculum (Treatment only)					
Simulation Embedded	10	0.65*	0.08	0.49	0.81
Stand Alone/related	18	0.52*	0.12	0.28	0.76
Stand Alone/not related	5	0.33*	0.15	0.05	0.62
NI	7	0.51*	0.20	0.11	0.89
Q-Between = 3.74, *df* = 2, *p* = .15 (Treatment only with NI excluded)					
Number of Sessions					
One Session	13	0.47*	0.11	0.25	0.69
Two to Three	7	0.44	0.21	0.03	0.84
Four - Six	8	0.76*	0.13	0.50	1.02
Eleven +	3	0.87*	0.33	0.22	1.52
NI	9	0.39*	0.12	0.15	0.62
Q-Between = 4.02, *df* = 3, *p* = .26 (With category NI excluded)					
Session Duration (In minutes)					
15-30 min.	4	0.22	0.17	-0.11	0.54
40-50 min.	11	0.67*	0.15	0.37	0.97
60-80 min	9	0.63*	0.11	0.41	0.85
90+ min.	3	0.54*	0.20	0.15	0.93
NI	13	0.46*	0.13	0.20	0.72
Q-Between = 5.13, *df* = 3, *p* = .16 (With category NI excluded)					
Total Duration (Constructed: # sessions x session time)					
Under 60 min.	6	0.38*	0.17	0.05	0.71
One to five hrs.	16	0.60*	0.11	0.38	0.81
Over 5 hrs.	5	0.68*	0.14	0.41	0.95
NI	13	0.46*	0.13	0.20	0.72
Q-Between = 1.91, *df* = 2, p = .38 (With category NI excluded)					

*$p < .05$

Scientific Inquiry and Reasoning Skills

(Q1 did not have enough studies, k = 6).

Demographic Moderator Variables for Research Question 2

Exhibit 23. Comparison of Levels of Demographic Moderator Variables for Research Question 2 for Scientific Inquiry and Reasoning Skills

Levels	Slope, Intercept and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Grade Ranges					
Grades 9-12	9	0.39*	0.18	0.03	0.75
Multiple Ranges	2	0.48	0.29	-0.09	1.06
Q-Between = 0.07, *df* = 1, *p* = .79					
Language of Instruction					
English	3	0.57*	0.23	0.12	1.02
Non-English	6	0.46*	0.15	0.18	0.75
NI	2	0.03	0.85	-1.63	1.69
Q-Between = 0.16, *df* = 1, *p* = .69 (With category NI excluded)					
Regions of the World					
U.S.	3	0.57*	0.23	0.12	1.02
Europe	2	-0.03	0.27	-1.55	0.50
Malaysia & Turkey	5	0.64*	0.12	0.41	0.87
NI	1	*g* = -0.82*	0.32	-1.44	-0.19
Q-Between = 5.19, *df* = 2, *p* = .08 (With category NI excluded)					

*$p < .05$

Substantive Moderator Variables for Research Question 2

Exhibit 24. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Scientific Inquiry and Reasoning Skills

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Simulation Type					
Phenomenon-based	8	0.55*	0.11	0.34	0.76
Virtual Labs	3	-0.06	0.53	-1.10	0.98
Q-Between = 1.26, *df* = 1, *p* = .26					
Nature of Modification to Treatment					
Representations	1	g = 0.88	0.37	0.16	1.59
Scaffolding	6	0.19	0.25	-0.30	0.68
Cooperative Learning	4	0.60*	0.13	0.36	0.85
Q-Between = 2.22, *df* = 41 *p* = .14 (Category "Representations" excluded)					

*$p < .05$

Exhibit 24. Comparison of Levels of Substantive Moderator Variables for Research Question 2 for Scientific Inquiry and Reasoning Skills (Continued)

Levels	Effect Size and Standard Error			Confidence Interval	
	k	g^+	*SE*	*Lower 95th*	*Upper 95th*
Assessment Delivery Mode					
Not Technology-based	3	0.67*	0.23	0.23	1.11
Technology-based Embedded	8	0.33	0.18	-0.01	0.68
Q-Between = 1.36, *df* = 1, *p* = .24					
Collaborative Settings (Treatment only; [1]Required *k* = 1)					
Collaboration Required	2	0.48	0.29	-0.09	1.06
Collaboration Optional	4	0.60*	0.13	0.36	0.85
No Collaboration	1	g = 0.18	0.37	-0.55	0.91
NI	4	0.17	0.44	-0.69	1.04
Q-Between = 0.97, *df* = 2, *p* = .62 (Treatment only with NI excluded)					
Flexibility (Treatment only)					
Free Form	1	g = -0.25	0.39	-1.07	0.52
Some Structure	6	0.32	0.23	-0.12	0.77
Very Structured	4	0.63*	0.19	0.26	1.00
Q-Between = 1.06, *df* = 1, p = .30 (Treatment only with "Free Form" excluded)					
Curriculum (Treatment only)					
Simulation Embedded	4	0.60*	0.13	0.36	0.85
Stand Alone/related	6	0.35*	0.39	-0.13	0.83
NI	1	g = -0.25*	0.25	-1.01	0.52
Q-Between = 0.83, *df* = 1, *p* = .36 (Treatment only with NI excluded)					
Number of Sessions					
One Session	2	-0.03	0.27	-0.55	0.50
Two - Three	2	0.03	0.85	-1.63	1.69
Four - Six	4	0.60*	0.13	0.36	0.85
Seven - Ten	1	g = 0.88	0.37	0.16	1.59
Eleven +	2	0.48	0.29	-0.09	1.06
Q-Between = 4.77, *df* = 3, *p* = .19 (With category "Seven - Ten" excluded)					
Session Duration (In minutes)					
40-50 min.	4	0.44	0.23	-0.01	0.89
60-80 min.	5	0.35	0.26	-0.17	0.86
90+	2	0.53	0.35	-0.16	1.22
Q-Between = 0.19, *df* = 2, *p* = .91					
Total Duration (Constructed: # sessions x session time)					
Under 60 min.	1	g = -0.25	0.39	-1.01	0.52
One to five hrs.	3	0.08	0.52	-0.94	1.09
Over 5 hrs.	7	0.58*	0.11	0.36	0.79
Q-Between = 0.89, *df* = 1 *p* = .35 (With category "Under 60 min." excluded)					

*$p < .05$

Assessment Information

There were 76 assessments listed across the 59 articles. Some articles contained multiple outcome measures which each corresponded to unique assessments. In other articles the same assessment was used for multiple studies leading to the same assessment corresponding to multiple effect sizes. Each assessment was coded in order to examine characteristics of the assessments such as: 1) when it was taken; 2) how many items it contained; 3) who developed it; and, 4) its validity and reliability.

Not all articles contained sufficient information on the administered assessments. For example, only 13 complete assessment instruments were included within their respective articles. Sample items were provided for 28 assessments, but there were 34 assessments identified in articles for which no items were included. (Note that three of the instruments were based on the analysis of student click-stream data (i.e., log files), and therefore no instrument could be provided.)

In order to obtain more information on assessments in the study, authors were contacted by email. Information was requested on the designers of the instrument, the constructs targeted by the instrument, the number and format of the items (or for a copy of the instrument), the reliability and validity of the instrument and the time period between the intervention and the administration of the instrument. Currently authors representing 19 articles of the 59 articles (32%) have responded with additional information. The information received from these authors was coded and is included in the analysis below.

The majority of the Q1 (simulation versus no simulation) studies used assessments that were not technology based, as the control group did not use technology. An exception was a study conducted by Weller (1995) that used a multiple choice assessment that was given on the computer. Another study embedded the assessment into the assignment (Michael, 2001). In this study the control group was using LEGO bricks, while the experimental group used a simulation involving LEGO bricks. In this study, students were asked to create a "creature" that would be found on a LEGO Planet (Michael, 2001). The "original" and "useful" sub-scales of the creative product semantic scale were used to judge the creativity of the students' artifacts.

There were 10 studies for the Q2 research question (modified simulations versus non-modified simulations) that included technology-based assessments. Five of these studies used embedded assessments. Hulshof and de Jong (2006), for example, examined the type of operations a student used while interacting in a geometry environment. Lohner, et al., (2005) used log files to obtain a measure of students' ability to model within different environments. Vreman-de Olde & de Jong (2006) used log files to determine how many design investigations students can create in their simulation environment. White and Frederiksen (1988) included two assessments that again examined students' work products while using the simulation. The remaining 5 assessments were combinations of multiple choice items and constructed response items.

Exhibit 25. The Frequency Count of the Delivery Mode of the Assessments (5 Assessments Were Used to Answer Both Q1 and Q2)

Delivery Mode	Frequency (all)	Frequency (Q1)	Frequency (Q2)
Not technology based	53	39	18
Embedded within the simulation	6	1	5
Technology based, but not embedded in the simulation	6	1	5
Not indicated	11	3	8

Most of the assessments were researcher designed (58 out of the 76 assessments, see Exhibit 26). Often researchers worked with teachers to create the assessment, or teachers reviewed the researcher-designed assessments. One study used a standardized assessment, the Scientific Assessment Test (Akcay, 2003). There were three assessments for which the designer of the assessment was not indicated and the source of the assessment was not identified.

As detailed in Exhibit 27, about one third of the assessments administered in the studies included in this meta-analysis were from previously developed assessment materials. Of these assessments only one research article did not state which source these previously developed materials were from. Approximately one third of the assessments were designed specifically for the research study discussed in the articles. The source for remaining one third was not identified.

Most of the assessments were based on multiple choice and/or constructed response items (see Exhibit 28). There were five assessments (discussed above) that used interactive computer tasks and nine assessments for which the articles did not indicate the type of items. For assessments of non-cognitive measures, most of the assessments were surveys using Likert scales (see Exhibit 28). The one interactive computing task (ICT) was a measure of creativity in the Michael (2001) study and for this measure only the experimental group (and not the control group) had a computer task as part of the assessment.

The number of items on these assessments varied (see Exhibit 29) ranging from 1 to 57 (although a few of the effect sizes were found by combining assessments and the total number of items used for the effect sizes is 104). The number of items also differed similarly across the content areas (see Exhibit 29). Note that while there are 5 ICTs only 2 of those used log files for analysis. The other three items were made up of constructed response and multiple choice items. In addition, while there were no assessments with a large number of items that were only constructed response items for all other item types there was a spread in the number of items on the assessments (see Exhibit 30).

As shown in Exhibit 31, most of the assessments were administered on the last day of the intervention or within two weeks after the intervention concluded. Only 5 were administered more than two weeks after the last day of the intervention. Of note is that there were 24 studies that did not indicate how much time had lapsed between the end of the intervention and the administration of the post assessment.

Reliability information was indicated for over half of the assessments. The majority reliability information was the report of Cronbach's alpha. Other reliability indicators included Cohen's Kappa and the Kuder Richardson coefficient (see Exhibit 32). Of the 29 assessments for which the reliability was not indicated, 6 of the assessments were directly taken from previous studies and therefore the reliability information might be discussed in other articles.

Over half of the assessments did not have information regarding validity (see Exhibit 33). Of these assessments, 17 came from previous studies where validity information might have been collected and/or reported. Of the assessments where the validity measure was indicated, the most common form of validity evidence was through an expert panel review. These experts were often K-12 teachers who had experience teaching the relevant subject matter.

Results from our analysis indicate that, when investigating the effectiveness of computer-based simulations in the classroom, most studies use researcher designed paper/pencil assessments. Moreover, we found that studies typically did not provide sufficient detail about the purposes, design method, and technical qualities of these assessments. While nearly a third of the assessments were taken from other studies, in those cases further information may be contained in the articles describing the previous study, most studies did not clearly describe the purpose and qualities of items or tasks that comprised the assessments used.

Exhibit 26. The Frequency Count by Assessment Designer Category (*N* = 76)

Assessment Designer	Frequency
Researcher	58
Standardized Measure	1
Not Indicated	17

Exhibit 27. The Frequency Count by Assessment Source (*N* = 76)

Assessment Source	Frequency
Based on a previously developed measure	27
Designed for the study	20
Not Indicated	29

Exhibit 28. The Frequency Count of the Type of Items by Assessment Construct (*N* = 76)

Item Type	Content knowledge	Reasoning Skills	Non-cognitive	Total
Multiple Choice	12	1	8	21
Constructed Response	10	1	0	11
Some multiple choice and some constructed response	27	3	0	30
ICT item	3	1	1	5
Not indicated	5	3	1	9

Exhibit 29. The Frequency Count of the Number of Items on the Assessment by Outcome Category (*N* = 76)

Number of items	Content knowledge	Reasoning	Non-Cognitive	Total
< 10	15	1	1	17
10 – 15	13	3	3	19
16 – 20	8	0	0	8
21 – 30	6	1	1	8
> 30	8	2	2	12
No items (log file analysis)	1	0	1	2
Not indicated	6	2	2	10

Exhibit 30. The Frequency Count of the Number of Items on the Assessment by Item Type (*N* = 76)

Number of items	MC	CR	MC and CR	ICT	NI
< 10	3	6	7	1	0
10 – 15	7	2	8	0	2
16 – 20	2	3	2	0	1
21 – 30	3	0	5	0	0
> 30	5	0	5	0	2
No items (log file analysis)	0	0	0	2	0
Not indicated	1	0	3	2	4

Exhibit 31. The Frequency Count of When the Assessments Were Delivered to the Student (*N* = 76)

Assessment Time Delivered	Frequency
Within a day of the intervention (includes embedded assessments)	29
Within two weeks of the intervention	18
More than two weeks after the intervention	5
Not Indicated	24

Exhibit 32. The Frequency Count of the Source of the Reliability Information (*N* = 76)

Reliability Measure	Frequency
Cronbach's alpha	31
Kuder-Richardson	5
Cohen's Kappa	3
Inter-rater correlation	2
Multiple methods	3
Other reliability information	3
Not indicated	29

Exhibit 33. The Frequency Count of the Source of the Validity Information (Out of 76)

Validity Measure	Frequency
Expert panel review	14
Statistical analysis of pilot test	8
Content validity from inter-rater reliability	4
Multiple validity information	2
Not indicated	48

Of note is that relatively few studies incorporated aspects of the simulation in their assessment design. Part of this may be due to the fact that for research question 1 the control group did not have access to the simulation, but another part may be due to the fact that analysis techniques such as log file analysis is still relatively new.

Reliability measures were reported more often than validity measures, with the most common reported measure of reliability being internal consistency. Validity was often established by having experts review the assessments to determine the content validity and representativeness of the assessment items. Some studies also included inter-rater reliability, and psychometric analyses performed using data collected from pilot studies as evidence of the validity of the inferences drawn from the assessment.

Overall, within the collection of studies that reported sufficient information on their assessments, we found wide variation in the types of constructs addressed and in the number and types of items included on the assessments. The studies also varied on the length of time between the use of the simulation and the administration of the assessment.

Discussion

In this meta-analysis, we closely examined studies that compared computer-based simulation learning conditions versus non-simulation learning conditions. We also examined studies that compared modified computer-based simulation learning conditions versus the same computer-based simulation without modification. Outcome measures of interest included achievement, scientific inquiry, and non-cognitive measures. A question that arose early in our work was whether we needed to analyze the three outcome measures we identified in the literature separately by research questions to perform the meta-analysis. We found that 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. Additionally, by not doing so, we would have increased the likelihood of conflating our multiple research questions and/or outcome measures and subsequently draw 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 these different groupings should not be directly compared.

Research Question 1 (Simulation versus no simulation)

Regarding our first research question (simulation versus no simulation), our results show that computer-based simulations have an advantage in achievement over non-simulation instruction. Many prior literature reviews (e.g., Clark, Nelson, Sengupta, & D'Angelo, 2009; Scalise et al., 2011; Smetana & Bell, 2012) have reached a similar conclusion. This meta-analysis, however, was able to quantify the magnitude of the average improvement due to simulations and examine specific moderator variables.

Of the 46 effect sizes for achievement only 4 of them had simulations in the subject area of mathematics. The average of these four effect sizes was negative and there was a statistically significant difference between these effect sizes and those for science. Therefore, these two subsets needed to be separated in subsequent moderator variable analyses. Since mathematics had only 4 effect sizes, moderator variable analyses were only performed on the science effect sizes. The moderator variable analysis showed that no significant differences existed across the K–12 age groups, language of instruction, or regions of the world where studies were conducted. Nor were significant differences found across different group sizes (individual vs. dyads vs. small groups), how flexible the simulation was, or the relationship of the simulation to the curriculum (whether it was embedded, related, or not directly related to the curriculum).

There was a significant difference found based on the number of sessions in which students used a simulation in science. Studies in which the simulation was used during four to six sessions had a higher effect size than studies that provided fewer or greater numbers of sessions. While it might be expected that studies with four to six sessions would have higher effect sizes than studies in which students were only exposed to the simulation one to three times (as students who have more exposure would presumably have more opportunity to learn the material), we also found that the studies in which students were exposed to the simulation more than 11 times had a lower effect size. One possible explanation for this is that for the eight studies in which there were more than 11 sessions students who did not use the simulation still might have had enough instruction so that they were able to learn the material that was reinforced in the simulation condition. The most common length of time for a session was between 40-50 minutes (i.e., one typical class period). No significant difference was found between the different session durations; however, this may be due to the low number of effect sizes in some of the categories.

Overall, our results indicate that using computer-based simulations in science, in many different configurations and contexts within the classroom, does improve student science achievement compared to not using simulations.

For scientific inquiry and reasoning skills, while the overall effect size was positive in favor of the simulation condition, the meta-analysis did not find a significant difference between the simulation condition and the non-simulation condition. This may be because only six effect sizes were included in the analysis and they had a wide standard error. Moderator variable analysis was not performed due to the small number of overall effect sizes.

There were 12 effect sizes in the non-cognitive measures outcome category. The average effect size was positive and it was found that there was a statistically significant difference in favor of the simulation condition. While the moderator variable analysis did find some statistically significant differences, the small sample size in each of the categories suggests that these differences should be interpreted with caution and these results should not be generalized.

Research Question 2 (Modified simulations versus non-modified simulations)

Regarding our second research question (modified simulations versus non-modified simulations), our results show that simulations supplemented or modified with some other form of instructional treatment (e.g., simulation plus scaffolding, simulation plus special representations) provided modest improvements in learning over non-modified simulations. For this research question there were 10 effect sizes in math and 40 effect sizes in science. While the average effect size for math was found to be positive, it was not found to be statistically significantly different from zero. Moderator variable analyses were run for both science and math combined. The two engineering effect sizes were included in the science sample as their average effect sizes and standard errors were similar.

Many different kinds of modifications or enhancements were used in the modified simulations. The types of modifications did cluster in a few general areas, specifically: scaffolding, representations, haptic feedback (feedback involving touch), the addition of a hands-on (real) laboratory activity, and cooperative learning. While most of these categories had too few effect sizes from which to make generalizations, the representations and scaffolding categories were both found to have a statistically significant average effect. Overall, there was not a statistically significant difference between the different types of modifications. For achievement outcomes in math and science, no significant differences were found across simulation type.

It is interesting to note that even when computer simulations were used in both conditions, most of the assessments administered were not technology-based, which indicates that while technology was used for learning, assessment still tends to be in the more traditional paper and pencil format. Another interesting result was that effect sizes associated with non-embedded technology-based assessments ¬– where the assessment technology was outside of or different than the learning technology – were not found to be significantly different from zero. There were only four effect sizes in this category and so not much can be generalized from this, but it would be interesting to investigate how the use of assessment technologies that are different from technology used for learning interacts with students' ability to demonstrate what they have learned.

Another variable for which there was a statistically significant difference between groups was the flexibility of the simulation. It was found that when the simulation was very structured the average effect size was higher than when the simulation allowed for more flexibility in the student interaction with the simulation. This is another a place in which more studies would be needed in order to further explain these results.

Other moderator variables such as whether or not the simulation required collaboration, whether or not students worked in groups, whether the simulation was embedded in the curriculum, how related to or separate from the curriculum the simulation was, and the number and duration of sessions were not found to be statistically significant from each other. The results of the moderator variable analysis for just the science effect sizes were consistent with the results when both the math and science effect sizes were used.

There were 11 effect sizes related to scientific inquiry and reasoning skills and 3 effect sizes related to non-cognitive outcomes for research question 2. For both of these outcome variables the average effect size was found to be in favor of the simulations with modifications and was statistically significant from zero. Since the non-cognitive outcome category had such a small sample size this result should be interpreted with caution and should not be generalized. Additional moderator analyses were not run for this group. While the moderator variable analysis was run for the scientific inquiry and reasoning skills outcome category, these results should also be interpreted with caution as the small sample size in the different groups makes these results not generalizable.

Limitations

It was noted that articles that contained detailed information about specific simulation features typically did not also include a research study and were therefore excluded from systematic review at either the abstract or article stage. On the other hand, articles with study details and outcome measures typically did not have as many details about the design or features of the simulation. Moreover, articles on some simulations that were commonly used (such as NetLogo) might not contain as many details as expected because they are well known. Consequently, articles that provided more detail about the features of the simulations studied were underrepresented in the meta-analysis. As a result, we could not conduct moderator analysis of specific simulation features.

Few research articles meeting our criteria in mathematics, technology, and engineering,[5] were identified. Engineering simulation studies can be identified by including college-age students in the sample; however, the same does not seem to be true of simulations in the domain of mathematics and technology. Overall, this suggests a need for high quality research studies that include simulation across a range of STEM disciplines in grades K-12.

Meta-analyses only include research studies that report an effect size or the data that allows the calculation of an effect size. This excludes studies that use only qualitative methods. Qualitative studies can explore research questions similar to those discussed in this meta-analysis, but also can address other related questions not addressed in this meta-analysis and suggest other avenues of research. While conducting the literature search for this meta-analysis, we identified 83 such studies that included computer-based simulations for STEM learning, but only collected qualitative data. Additionally, some of the quantitative articles that were included in the meta-analysis also contained additional research results that came from qualitative methods that could not be reported on here. These two pools of articles include a vast amount of evidence that could be used to better understand and support the findings reported in this study.

5 There were many studies involving engineering education, but they were excluded because they used college-age or older students.

Conclusion

In this report, we have described our systematic review and meta-analysis of the literature on computer simulations designed to support science, technology, engineering, and mathematics (STEM) learning in K-12 instructional settings. Both quantitative and qualitative research studies on the effects of simulation in STEM were reviewed. Only studies that reported effect size measures or the data necessary to calculate effect sizes were included in the meta-analysis. Initial search results identified 260 articles, of which 59 (23%) met this requirement. Results from the meta-analysis of 59 studies indicate that, overall, simulations have a beneficial effect over treatments in which there were no simulations. Also, simulations with modifications were shown to have a beneficial effect over simulations without those modifications. Most of the assessments used to measure outcomes associated with simulations were paper/pencil based; few took advantage of the affordances of the technology involved in the simulations that were studied. It is important to note that the studies included in the meta-analysis were predominately in science education, suggesting that an important need is a more robust pool of high quality research studies on simulations in other STEM domains at the K-12 level. Thus, while our work shows that simulations, in many different configurations or contexts within the classroom, can improve student learning, there is still much to be learned about the educational benefits of computer simulations across the STEM domains.

References

Abrami, P. C., & Bernard, R. M. (2013). Statistical control versus classification of study quality in meta-analysis. *Effective Education, 4*(1), 43-72. doi: 10.1080/19415532.2012.761889

Akçay, H., Feyzioglu, B., & Tüysüz, C. (2003). Kimya ögregtiminde Bilgisayar Benzesimlerinin Kullaniminin Lise Ögrencilerinin Bagarisina ve Tutumuna Etkisi. *Kuram ve Uygulamada Egitim Bilimleri, 3*(1), 7–26.

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.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2011). *Comprehensive meta-analysis version* 2.2.064. Englewood, NJ: Biostat.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

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.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel Plot–Based Method of Testing and Adjusting for Publication Bias in Meta Analysis. *Biometrics, 56*(2), 455-463.

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.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). *Measuring inconsistency in meta-analyses. British Medical Journal, 327*(7414), 557–560. doi: 10.1136/bmj.327.7414.557

Hulshof, C. D., & De Jong, T. (2006). Using just-in-time information to support scientific discovery learning in a computer-based simulation. *Interactive Learning Environments, 14*(1), 79–94. doi:10.1080/10494820600769171

Löhner, S., Van Joolingen, W. R., Savelsbergh, E. R., & Van Hout-Wolters, B. (2005). Students' reasoning during modeling in an inquiry learning environment. *Computers in Human Behavior, 21*(3 SPEC. ISS.), 441–461.

Michael, K. Y. (2001). The Effect of a Computer Simulation Activity versus a Hands-on Activity on Product Creativity in Technology Education. *Journal of Technology Education, 13*(1), 31–43.

Ozgun-Koca, S. A. (2004). The Effects of Multiple Linked Representations on Students' Learning of Linear Relationships. *Hacettepe University Journal of Education, 26*, 82–90.

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.

Rothstein, H., Sutton, A., & Borenstein, M. (2005). *Publication bias in meta-analysis*. West Sussex, England: John Wiley and Sons.

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.

Shadish, W. R., Cook, T., & Campbell, D. T. (2001). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

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.

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13,* 130-149.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112-125.

Vreman-de Olde, C., & De Jong, T. (2006). Scaffolding Learners in Designing Investigation Assignments for a Computer Simulation. *Journal of Computer Assisted Learning, 22*(1), 63–73.

Weller, H. G. (1995). Diagnosing and Altering Three Aristotelian Alternative Conceptions in Dynamics: Microcomputer Simulations of Scientific Models. J*ournal of Research in Science Teaching, 32*(3), 271–290.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction, 16*(1), 3–118. doi:10.1207/s1532690xci1601_2

Appendix: Citations of Articles Included in the Meta-Analysis

Abdullah, S., & Shariff, A. (2008). The effects of inquiry-based computer simulation with cooperative learning on scientific thinking and conceptual understanding of gas laws. *Eurasia Journal of Mathematics, Science and Technology Education, 4(*4), 387–398.

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the Effects of Different Multiple Representational Systems in Learning Primary Mathematics. *Journal of the Learning Sciences, 11*(1), 25–61.

Akçay, H., Feyzioglu, B., & Tüysüz, C. (2003). Kimya ögregtiminde Bilgisayar Benzesimlerinin Kullaniminin Lise Ögrencilerinin Bagarisina ve Tutumuna Etkisi. *Kuram ve Uygulamada Egitim Bilimleri, 3*(1), 7–26.

Akpan, J. P., & Andre, T. (2000). Using a computer simulation before dissection to help students learn anatomy. *Journal of Computers in Mathematics and Science Teaching, 19(*3), 297–313.

Akpan, J., & Strayer, J. (2010). Which Comes First: The Use of Computer Simulation of Frog Dissection or Conventional Dissection as Academic Exercise? *Journal of Computers in Mathematics and Science Teaching, 29*(2), 113–138.

Ardac, D., & Sezen, A. H. (2002). Effectiveness of Computer-Based Chemistry Instruction in Enhancing the Learning of Content and Variable Control Under Guided Versus Unguided Conditions. *Journal of Science Education and Technology, 11*(1), 39–48. doi:10.1023/A:1013995314094

Barnea, N., & Dori, Y. J. (1999). High-School Chemistry Students' Performance and Gender Differences in a *Computerized Molecular Modeling Learning Environment. Journal of Science Education and Technology, 8*(4), 257–271.

Baxter, J. H., & Preece, P. F. W. (1999). Interactive multimedia and concrete three-dimensional modelling. *Journal of Computer Assisted Learning, 15*(4), 323–331. doi:10.1046/j.1365-2729.1999.00107.x

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence, 19*(3-4), 363–392. doi:10.1080/08839510590910200

Chang, K.-E., Chen, Y.-L., Lin, H.-Y., & Sung, Y.-T. (2008). Effects of Learning Support in Simulation-Based Physics Learning. *Computers & Education, 51*(4), 1486–1498.

Çigrik, E., & Ergül, R. (2009). The investigation of the effect of simulation based teaching on the student achievement and attitude in electrostatic induction. *Procedia - Social and Behavioral Sciences, 1*(1), 2470–2474.

Clark, D., & Jorde, D. (2004). Helping Students Revise Disruptive Experientially Supported Ideas about Thermodynamics: Computer Visualizations and Tactile Models. *Journal of Research in Science Teaching, 41(*1), 1–23.

Eglash, R., Krishnamoorthy, M., Sanchez, J., & Woodbridge, A. (2011). Fractal Simulations of African Design in Pre-College Computing Education. *ACM Transactions on Computing Education, 11*(3).

Eskrootchi, R., & Oskrochi, G. R. (2010). A Study of the Efficacy of Project-Based Learning Integrated with Computer-Based Simulation--STELLA. *Educational Technology & Society, 13*(1), 236–245.

Eylon, B.-S., & Others, A. (1996). Computer Simulations as Tools for Teaching and Learning: Using a Simulation Environment in Optics. *Journal of Science Education and Technology, 5*(2), 93–110.

Faulkner, D., Joiner, R., Littleton, K., Miell, D., & Thompson, L. (2000). The mediating effect of task presentation on collaboration and children's acquisition of scientific reasoning. European *Journal of Psychology of Education, 15*(4), 417–430. doi:10.1007/BF03172985

Frederiksen, J. R., White, B. Y., & Gutwill, J. (1999). Dynamic mental models in learning science: The importance of constructing derivational linkages among models. *Journal of Research in Science Teaching, 36*(7), 806–836. doi:10.1002/(SICI)1098-2736(199909)36:7<806::AID-TEA5>3.0.CO;2-2

Friedler, Y., & Others, A. (1992). Problem-Solving Inquiry-Oriented Biology Tasks Integrating Practical Laboratory and Computer. *Journal of Computers in Mathematics and Science Teaching, 11*(3), 347–357.

Fund, Z. (2007). The effects of scaffolded computerized science problem-solving on achievement outcomes: A comparative study of support programs. *Journal of Computer Assisted Learning, 23*(5), 410–424. doi:10.1111/j.1365-2729.2007.00226.x

Geban, Ö., Askar, P., & Özkan, Ï. (1992). Effects of computer simulations and problem-solving approaches on high school students. *The Journal of Educational Research, 86*(1), 5–10. doi:10.1080/00220671.1992.9941821

Gelbart, H., Brill, G., & Yarden, A. (2009). The Impact of a Web-Based Research Simulation in Bioinformatics on Students' Understanding of Genetics. *Research in Science Education, 39*(5), 725–751.

Gutwill, J. P., Frederiksen, J. R., & White, B. Y. (1999). Making Their Own Connections: Students' Understanding of Multiple Models in Basic Electricity. *Cognition and Instruction, 17*(3), 249–282. doi:10.1207/S1532690XCI1703_2

Han, I., & Black, J. B. (2011). Incorporating Haptic Feedback in Simulation for Learning Physics. *Computers & Education, 57*(4), 2281–2290.

Hulshof, C. D., & De Jong, T. (2006). Using just-in-time information to support scientific discovery learning in a computer-based simulation. *Interactive Learning Environments, 14*(1), 79–94. doi:10.1080/10494820600769171

Huppert, J., Lomask, S. M., & Lazarowitz, R. (2002). Computer Simulations in the High School: Students' Cognitive Stages, Science Process Skills and Academic Achievement in Microbiology. *International Journal of Science Education, 24*(8), 803–821.

Hurst, R. W., & Milkent, M. M. (1996). Facilitating successful prediction problem solving in biology through application of skill theory. *Journal of Research in Science Teaching, 33*(5), 541–552. doi:10.1002/(SICI)1098-2736(199605)33:5<541::AID-TEA5>3.0.CO;2-R

Ioannidou, A., Repenning, A., Webb, D., Keyser, D., Luhn, L., & Daetwyler, C. (2010). Mr. Vetro: A collective simulation for teaching health science. *International Journal of Computer-Supported Collaborative Learning, 5*(2), 141–166. doi:10.1007/s11412-010-9082-8

Jaakkola, T, & Nurmi, S. (2008). Fostering elementary school students' understanding of simple electricity by combining simulation and laboratory activities. *Journal of Computer Assisted Learning, 24*(4), 271–283. doi:10.1111/j.1365-2729.2007.00259.x

Jaakkola, Tomi, Nurmi, S., & Veermans, K. (2011). A comparison of students' conceptual understanding of electric circuits in simulation only and simulation-laboratory contexts. *Journal of Research in Science Teaching, 48*(1), 71–93. doi:10.1002/tea.20386

Johnson-Glenberg, M. C., Birchfield, D., & Usyal, S. (2009). "SMALLab": Virtual Geology Studies Using Embodied Learning with Motion, Sound, and Graphics. *Educational Media International, 46*(4), 267–280.

Kinzie, M. B., Strauss, R., & Foss, M. J. (1993). The effects of an interactive dissection simulation on the performance and achievement of high school students. *Journal of Research in Science Teaching, 30*(8), 989–1000. doi:10.1002/tea.3660300813

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching, 44*(1), 183–203. doi:10.1002/tea.20152

Kolloffel, B., Eysink, T. H. S., & De Jong, T. (2011). Comparing the effects of representational tools in collaborative and individual inquiry learning. *International Journal of Computer-Supported Collaborative Learning, 6*(2), 223–251. doi:10.1007/s11412-011-9110-3

Lalley, J. P., Piotrowski, P. S., Battaglia, B., Brophy, K., & Chugh, K. (2010). A Comparison of V-Frog[C] to Physical Frog Dissection. *International Journal of Environmental and Science Education, 5*(2), 189–200.

Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Annual Meeting of the American Educational Research Association,* April 2006, San Francisco, US; Parts of this study were presented at the aforementioned meeting., 98(4), 902–913. doi:10.1037/0022-0663.98.4.902

Leelawong, K., & Biswas, G. (2008). Designing Learning by Teaching Agents: The Betty's Brain System. International *Journal of Artificial Intelligence in Education, 18*(3), 181–208. Retrieved from http://dl.acm.org/citation.cfm?id=1454278.1454280

Liu, H., & Su, I. (2011). Learning residential electrical wiring through computer simulation: The impact of computer-based learning environments on student achievement and cognitive load. *British Journal of Educational Technology, 42*(4), 598–607. doi:10.1111/j.1467-8535.2009.01047.x

Liu, H.-C., & Chuang, H.-H. (2011). Investigation of the Impact of Two Verbal Instruction Formats and Prior Knowledge on Student Learning in a Simulation-Based Learning Environment. *Interactive Learning Environments, 19*(4), 433–446.

Löhner, S., Van Joolingen, W. R., Savelsbergh, E. R., & Van Hout-Wolters, B. (2005). Students' reasoning during modeling in an inquiry learning environment. *Computers in Human Behavior, 21*(3 SPEC. ISS.), 441–461.

Michael, K. Y. (2001). The Effect of a Computer Simulation Activity versus a Hands-on Activity on Product Creativity in Technology Education. *Journal of Technology Education, 13*(1), 31–43.

Moreno, R., Reislein, M., & Ozogul, G. (2010). Using virtual peers to guide visual attention during learning: A test of the persona hypothesis. *Journal of Media Psychology: Theories, Methods, and Applications, 22*(2), 52–60. doi:10.1027/1864-1105/a000008

Ozgun-Koca, S. A. (2004). The Effects of Multiple Linked Representations on Students' Learning of Linear Relationships. *Hacettepe University Journal of Education, 26*, 82–90.

Papaevripidou, M., Constantinou, C. P., & Zacharia, Z. C. (2007). Modeling complex marine ecosystems: An investigation of two teaching approaches with fifth graders. *Journal of Computer Assisted Learning, 23*(2), 145–157. doi:10.1111/j.1365-2729.2006.00217.x

Pedaste, M., Sarapuu, T., & Education, O. of. (2006). The Factors Influencing the Outcome of Solving Story Problems in a Web-Based Learning Environment. *Interactive Learning Environments, 14*(2), 153–176.

Plass, J. L., Homer, B. D., Milne, C., Jordan, T., Kalyuga, S., Kim, M., & Lee, H. (2009). Design factors for effective science simulations: Representation of information. International *Journal of Gaming and Computer-Mediated Simulations, 1*(1), 16–35.

Riess, W., & Mischo, C. (2010). Promoting systems thinking through biology lessons. *International Journal of Science Education, 32*(6), 705–725.

Sierra-Fernandez, J. L., & Perales-Palacios, F. J. (2003). The effect of instruction with computer simulation as a research tool on open-ended problem-solving in a Spanish classroom of 16-year-olds. *Journal of Computers in Mathematics and Science Teaching, 22*(2), 119–140.

Strauss, R. T., & Kinzie, M. B. (1994). Student Achievement and Attitudes in a Pilot Study Comparing an Interactive Videodisc Simulation to Conventional Dissection. *American Biology Teacher, 56*(7), 398–402.

Suh, J., & Moyer, P. S. (2007). Developing Students' Representational Fluency Using Virtual and Physical Algebra Balances. *Journal of Computers in Mathematics and Science Teaching, 26*(2), 155–173.

Sun, K., Lin, Y., & Yu, C. (2008). A study on learning effect among different learning styles in a Web-based lab of science for elementary school students. *Computers & Education, 50*(4), 1411–1422. doi:10.1016/j.compedu.2007.01.003

Swaak, J., De Jong, T., & Van Joolingen, W. R. (2004). The effects of discovery learning and expository instruction on the acquisition of definitional and intuitive knowledge. *Journal of Computer Assisted Learning, 20*(4), 225–234. doi:10.1111/j.1365-2729.2004.00092.x

Trey, L., & Khan, S. (2008). How science students can learn about unobservable phenomena using computer-based analogies. *Computers & Education, 51*(2), 519–529. doi:10.1016/j.compedu.2007.05.019

Van der Meij, J, & De Jong, T. (2011). The Effects of Directive Self-Explanation Prompts to Support Active Processing of Multiple Representations in a Simulation-Based Learning Environment. *Journal of Computer Assisted Learning, 27*(5), 411–423.

Van der Meij, Jan, & De Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction, 16(*3), 199–212. doi:10.1016/j.learninstruc.2006.03.007

Veermans, K., Van Joolingen, W., & De Jong, T. (2006). Use of Heuristics to Facilitate Scientific Discovery Learning in a Simulation Learning Environment in a Physics Domain. *International Journal of Science Education, 28*(4), 341–361.

Vreman-de Olde, C., & De Jong, T. (2006). Scaffolding Learners in Designing Investigation Assignments for a Computer Simulation. *Journal of Computer Assisted Learning, 22*(1), 63–73.

Weller, H. G. (1995). Diagnosing and Altering Three Aristotelian Alternative Conceptions in Dynamics: Microcomputer Simulations of Scientific Models. *Journal of Research in Science Teaching, 32*(3), 271–290.

White, B. Y. (1993). ThinkerTools: Causal Models, Conceptual Change, and Science Education. *Cognition and Instruction, 10*(1), 1–100. doi:10.1207/s1532690xci1001_1

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction, 16(*1), 3–118. doi:10.1207/s1532690xci1601_2

Appendix B: Codebook for Simulation Meta-Analysis

Overview

We used FileMaker during article coding (see Appendix C for a screenshot of the coding interface). Each article had one record in FileMaker for each study comparison (i.e., effect size) of interest. If an article reported only one group comparison on one outcome, for example, that article had one record. If, on the other hand, an article reported one group comparison on two separate outcomes (e.g., a measure of content knowledge and a measure of engagement), there would be two records for the article (designated Comparison 1 and 2). The separate Comparison Effect Size (Excel) Spreadsheet document [not included in this report] that was prepared by part of the research team noted which comparisons were of interest for each study. Additionally, two researchers coded each article comparison for purposes of inter-rater reliability (thus, there will be a double set of records in FileMaker for each article).

Each record is divided into an upper and lower section (see Appendix C for a screenshot of the FileMaker layout). The upper section requires coding fields for the meta-data associated with each article. The metadata included: Study Identifier (AuthorYear), Authors, Title, Year), the comparison number, the coder's name, and other study characteristics that are applicable to the entire article (e.g, Research Design, Grade Range). Three other fields in the upper section pertain to the nature of the outcome variable and, thus, varied by comparison. The lower section of the record is displayed as two columns of vertical code fields, one column for codes relating to the Experimental Condition, and one column for codes relating to the Control Condition. At the bottom of the record page, there is a box available for typing in "Notes."

The coding fields outlined in red on the FileMaker record were taken directly from the separate Comparison Effect Size Spreadsheet document. There are drop-down menus for nearly all of the coding fields. Every field should have a code in it. In some cases, the required information may not be given in the article and one of two codes can be used: NI – the information is not indicated in the article, or NA – the information is not applicable. The difference between NI and NA is that for NI, you believe there should be a code for the field, but you can't tell from the article what it should be; for NA, you believe this field does not apply to the study.

The following table describes the coding fields and options for the FileMaker record page. The research team iterated on developing the codes, code options, and code definitions. They used this table as an aid during coding.

Exhibit B1. Codes

Coding Fields	Response Options	Explanation
	Upper Section Coding Fields	
Identifier	[open-ended]	Identifies the article being coded in the format Author(s) LastName_Year
Author	[open-ended]	Full list of authors on the article being coded (APA format)
Title	[open-ended]	Full title of the article
Year	[open-ended]	Year the article was published
Coder	[open-ended]	Name of the person coding
Comparison	[open-ended]	The number corresponding to the comparison (or effect size) being coded
Research Question	no sim v sim sim v sim + modifications	**No sim v sim** - The study compares a simulation to a condition not involving a simulation, where the latter is the Control Group **Sim v sim + modifications** – The study compares the simulation to a condition where individuals receive a modified version of the simulation or setting; in this case, the simulation with modification is the Experimental Group.
Research Design	1 - RCT 2 - quasi-experimental 3 - pre-experimental 4 - qualitative	This information can be inferred from Column S of the Comparison Effect Size Spreadsheet. **RCT (randomly controlled trial) –** an experimental study that uses random assignment of individuals to the Experimental and Control groups **Quasi-experimental (QE) –** an experimental study that does not use random assignment of individuals (e.g., assignment of intact classes to Experimental and Control groups) **Pre-Experimental –** the study does not include a Control Group (e.g., one group pre-test post-test design) **Qualitative –** the study uses descriptive information and does not manipulate the independent variables
Outcome Measure	content knowledge/achievement subject matter attitude technology attitude scientific inquiry reasoning skills other	Focus of the outcome measure for the comparison.

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
	Upper Section Coding Fields	
Group Equiv Method	true individual random assignment block randomization pretest difference test pretest (groups within .25 SD) covariate adjusted pretest (b/t .25 and .50 SD) participant matching NI *NA*	The method by which equivalency was established between the Experimental and Control Groups.
Attrition in Treatment	Negligible violates research design assumptions NI	The extent to which there was attrition by participants in either or both the Experimental and Control Groups. Significant attrition may indicate that the research design assumptions have been violated.
Outcome Measure Source	1 – one-shot cumulative 2 – reported composite 3 – calculated composite (avg) 4 – selected individual item NA	The way in which the outcome measure for the comparison was calculated.
ES Extraction Procedure	0 – calculated from descriptive stats 1 – calculated from inferential stats 2 – estimated from reported p-values 3 – estimated with assumptions 4 – reported NA	The way in which the effect size was calculated from the data provided for the comparison.
Overall finding	experiment > control experiment < control no difference NI	**Experiment > control** – the study reports that the Experimental Group performed significantly better than the Control Group **Experiment < control** – the study reports that the Control Group performed significantly better than the Experimental Group **No difference** – the study reports that there was no significant difference between the Experimental and Control Groups **NI** – the finding is unclear

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
	Upper Section Coding Fields	
Pre/Post	post only pre and post delayed post only pre and delayed post post and delayed post pre, post, and delayed post other	The combination of pre- and post-testing used for the measure
Instructor Equivalence	Yes No NI	Whether the Experimental and Control Groups had the same or similar instructors
Material Equivalence	Yes No NI	Whether the Experimental and Control Group interventions were similar in the content covered
Time on Task Equivalence	Yes No NI	Whether the Experimental and Control Groups had a comparable time on task
Language	*Free Response* NI	Language in which the intervention was given.
Grade Range	K-5 6-8 9-12 multiple ranges	Grade range of subjects for the study comparison. In countries that use different grade level designations, the translation should be made to US grade equivalents
Location	*Free Response* NI	Country and/or region/city where the study took place
STEM Domain	Science Technology Engineering Mathematics other	Predominant content domain of the intervention
Learning Theory	[open-ended]	What learning theory or theories or assumptions were made in the development of the simulation? (only if explicitly mentioned in the article)

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
Lower Section Coding Fields **This section: Only fill out if RCT or Quasi-Experimental is checked – there will be one column for the experimental group and one for the control group**		
Experimental (and Control) Group	*Free Response*	Brief descriptions of the Experimental and Control Groups, taken from the Comparison Effect Size Spreadsheet (Columns G and H)
N E and N C	*Free Response*	Numbers for the Experimental and Control Groups (Columns N and O of the Comparison Effect Size Spreadsheet)
ES d	*Free Response*	The effect size for the comparison (Column Q of the Comparison Effect Size Spreadsheet)
Number of Sessions E (and C)	*Free Response* NI	How many sessions did the intervention last for the Experimental and Control Groups
Session Time E (and C)	*Free Response* NI	Time (in minutes) of each session
Duration of Intervention E (and C)	*Free Response* NI	The time frame from the start to the end of the intervention in days (e.g., 15 days) for the Experimental and Control Groups. It may be possible to estimate days from the session times. Do not include days in the study that were devoted only to testing.
Sim Topic E (and C)	*Free Response* NI	The specific topic of the simulation or Control intervention (e.g., "frog dissection and anatomy")
Sim Name E (and C)	*Free Response* NI	The formal or commercial name of the simulation or Control intervention
Sim Type E (and C)	virtual lab virtual world phenomenon sim agent-based Other NI NA	The dominant style of the simulation: **Virtual lab –** the setting is a virtual lab **Virtual world –** the setting is a virtual world **Phenomenon sim –** the simulation investigates a particular phenomenon and the setting isn't specified **Agent based –** the student is expected to interact with a virtual character

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
Lower Section Coding Fields **This section: Only fill out if RCT or Quasi-Experimental is checked – there will be one column for the experimental group and one for the control group**		
Experimental (and Control) Group	*Free Response*	Brief descriptions of the Experimental and Control Groups, taken from the Comparison Effect Size Spreadsheet (Columns G and H)
N E and N C	*Free Response*	Numbers for the Experimental and Control Groups (Columns N and O of the Comparison Effect Size Spreadsheet)
ES d	*Free Response*	The effect size for the comparison (Column Q of the Comparison Effect Size Spreadsheet)
Number of Sessions E (and C)	*Free Response* NI	How many sessions did the intervention last for the Experimental and Control Groups
Session Time E (and C)	*Free Response* NI	Time (in minutes) of each session
Duration of Intervention E (and C)	*Free Response* NI	The time frame from the start to the end of the intervention in days (e.g., 15 days) for the Experimental and Control Groups. It may be possible to estimate days from the session times. Do not include days in the study that were devoted only to testing.
Sim Topic E (and C)	*Free Response* NI	The specific topic of the simulation or Control intervention (e.g., "frog dissection and anatomy")
Sim Name E (and C)	*Free Response* NI	The formal or commercial name of the simulation or Control intervention
Sim Type E (and C)	virtual lab virtual world phenomenon sim agent-based Other NI NA	The dominant style of the simulation: **Virtual lab –** the setting is a virtual lab **Virtual world –** the setting is a virtual world **Phenomenon sim –** the simulation investigates a particular phenomenon and the setting isn't specified **Agent based –** the student is expected to interact with a virtual character

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
Lower Section Coding Fields This section: Only fill out if RCT or Quasi-Experimental is checked – there will be one column for the experimental group and one for the control group		
Collaborative E (and C)	requires collaboration collaboration optional not collaborative NI	The amount of collaboration required by the simulation (or Control intervention) – namely, how the simulation/ intervention is designed for use (but not necessarily how it was used in the study).
Flexibility E (and C)	very structured use some structured use free form use NI	How much flexibility the simulation (or Control intervention) allows the student – namely, how the simulation/intervention is designed for use (but not necessarily how it was used in the study)
Platform E (and C)	laptop/desktop tablet handheld calculator other tech not tech NI	Type of platform the simulation software (or Control intervention) runs on
Instructional Setting E (and C)	classroom after school program out of school informal learning setting NI	Where the Experimental and Control Group interventions took place
Curriculum E (and C)	embedded in curriculum stand-alone but related stand-alone but not related NI NA	The relationship between the simulation (or Control intervention) and the regular curriculum **Embedded in curriculum –** simulation/ intervention is closely integrated into the normal curriculum **Stand-alone but related –** simulation/ intervention is related to the current course curriculum but not tied directly to it **Stand-alone but not related -** simulation/ intervention is not related to the current course curriculum **NA –** simulation/intervention is out of school or in an informal learning setting

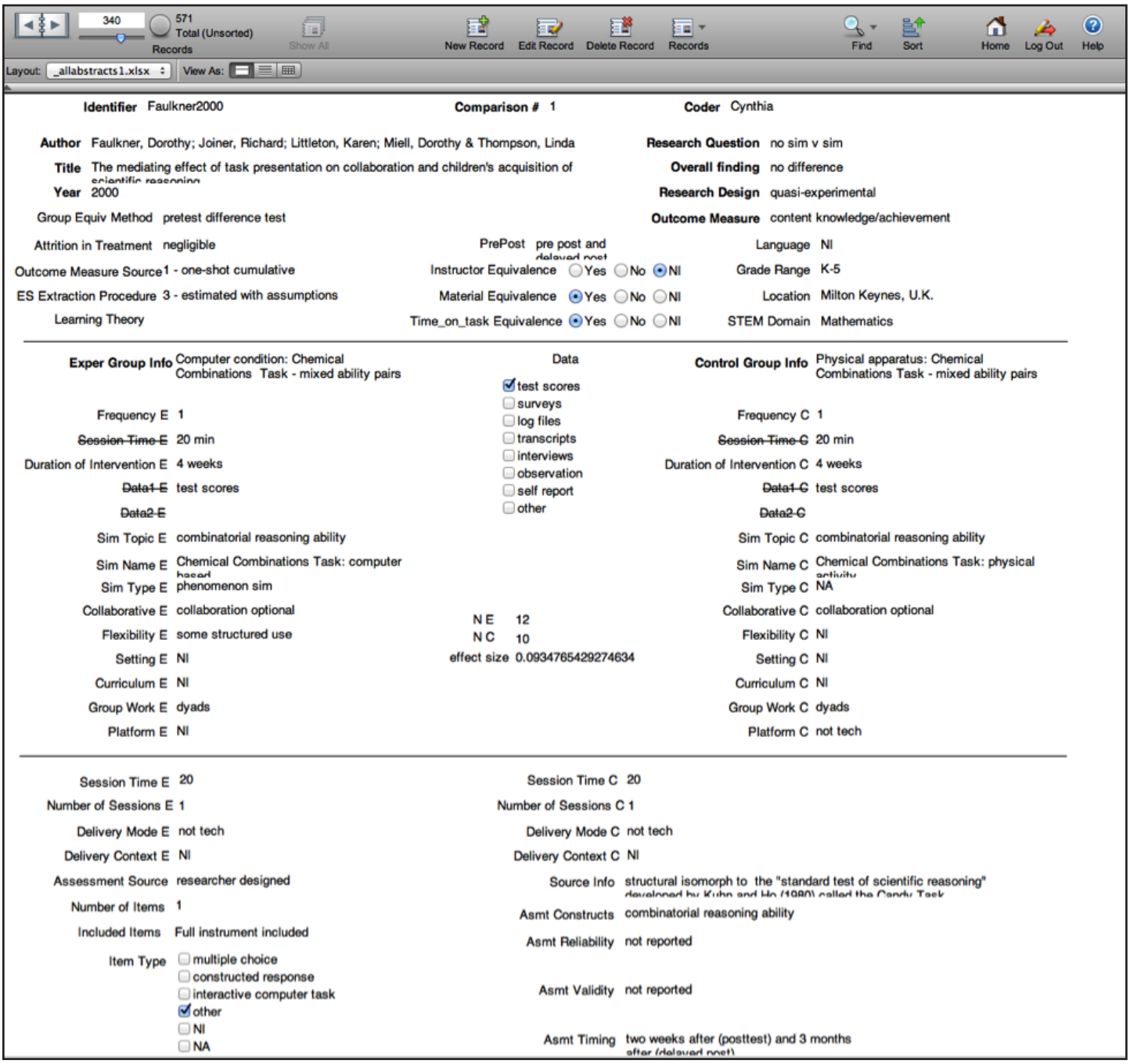
Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
Lower Section Coding Fields This section: Only fill out if RCT or Quasi-Experimental is checked – there will be one column for the experimental group and one for the control group		
Group Work E (and C)	individual work dyads small groups whole class NI	Way in which the students participated in the simulation or intervention: **Individual work –** used by individuals **Dyads –** used by pairs **Small Groups –** used by small groups **Whole class –** used by (or presented to) the whole class
Asmt Delivery Mode E (or C)	embedded tech, not embedded not tech NI NA	**Embedded –** the assessment is embedded within the simulation **Tech, not embedded –** the assessment is not embedded in the simulation but is delivered using technology **Not tech –** the assessment is not delivered using technology (e.g., paper and pencil)
Asmt Delivery Context E (or C)	instructional setting out of instructional setting NI NA	**Instructional setting –** assessment given in a classroom-type setting (whether or not it is during or outside the class period) **Out of instructional setting –** assessment given to students to do on their own
Lower Section Coding Fields This section: Applies to both the experimental and the control groups		
Data Source	test scores surveys log files transcripts interviews observation self-report other	The type of instrument used to gather the outcome data. If the outcome measure is a composite score (see Outcome Measure Source field). Coded as separate check boxes
Assessment Source	teacher designed researcher designed curriculum test district, state, or national test other standardized test NI NA	**Teacher designed –** classroom assessment designed by teacher (versus researcher) **Researcher designed –** assessment designed by researchers **Curriculum test –** assessment designed for the curriculum **District, state, or national test –** assessment given at the district, state, or national level **Other standardized test**

Exhibit B1. Codes (Continued)

Coding Fields	Response Options	Explanation
	Lower Section Coding Fields **This section: Applies to both the experimental and the control groups**	
Source Info	*Free Response* NI NA	A place to put additional notes about the designer(s) of the instrument or the origins of the items on the instrument
Assessment Construct(s)	*Free Response* NI NA	Constructs targeted by the instrument (specific domain/topic information)
Assessment: Total number of items	*Free Response* NI NA	This should be a number indicating the total number of items on the assessment
Assessment: Type of items	Multiple choice Constructed response Interactive computer tasks Other NI NA	The type of items
Included Items	No items included Some sample items Full instrument included NA	This indicates whether any or all of the instrument items are included in the article.
Assessment: Reliability Information	*Free Response* NI NA	This should be information about the type of reliability information reported, including the reliability coefficient.
Assessment: Validity Information	*Free Response* NI NA	This should be information on the validity measures for the assessment
Assessment Timing	*Free Response* NI NA	This indicates the time period between the intervention and the administration of the assessment
Notes	[open-ended]	Any additional notes about the article

Appendix C: FileMaker screenshot



340 571 Total (Unsorted) Records — Show All — New Record — Edit Record — Delete Record — Records — Find — Sort — Home — Log Out — Help

Layout: _allabstracts1.xlsx — View As:

Identifier Faulkner2000 **Comparison #** 1 **Coder** Cynthia

Author Faulkner, Dorothy; Joiner, Richard; Littleton, Karen; Miell, Dorothy & Thompson, Linda
Title The mediating effect of task presentation on collaboration and children's acquisition of scientific reasoning
Year 2000
Group Equiv Method pretest difference test
Attrition in Treatment negligible
Outcome Measure Source 1 - one-shot cumulative
ES Extraction Procedure 3 - estimated with assumptions
Learning Theory

Research Question no sim v sim
Overall finding no difference
Research Design quasi-experimental
Outcome Measure content knowledge/achievement

PrePost pre post and delayed post
Instructor Equivalence ○Yes ○No ◉NI
Material Equivalence ◉Yes ○No ○NI
Time_on_task Equivalence ◉Yes ○No ○NI

Language NI
Grade Range K-5
Location Milton Keynes, U.K.
STEM Domain Mathematics

	Exper Group Info	Control Group Info
	Computer condition: Chemical Combinations Task - mixed ability pairs	Physical apparatus: Chemical Combinations Task - mixed ability pairs
Frequency	1	1
~~Session Time~~	20 min	20 min
Duration of Intervention	4 weeks	4 weeks
~~Data1~~	test scores	test scores
~~Data2~~		
Sim Topic	combinatorial reasoning ability	combinatorial reasoning ability
Sim Name	Chemical Combinations Task: computer based	Chemical Combinations Task: physical activity
Sim Type	phenomenon sim	NA
Collaborative	collaboration optional	collaboration optional
Flexibility	some structured use	NI
Setting	NI	NI
Curriculum	NI	NI
Group Work	dyads	dyads
Platform	NI	not tech

Data
☑ test scores
☐ surveys
☐ log files
☐ transcripts
☐ interviews
☐ observation
☐ self report
☐ other

N E 12
N C 10
effect size 0.0934765429274634

Session Time E	20
Number of Sessions E	1
Delivery Mode E	not tech
Delivery Context E	NI
Assessment Source	researcher designed
Number of Items	1
Included Items	Full instrument included

Item Type
☐ multiple choice
☐ constructed response
☐ interactive computer task
☑ other
☐ NI
☐ NA

Session Time C	20
Number of Sessions C	1
Delivery Mode C	not tech
Delivery Context C	NI
Source Info	structural isomorph to the "standard test of scientific reasoning" developed by Kuhn and Ho (1980) called the Candy Task
Asmt Constructs	combinatorial reasoning ability
Asmt Reliability	not reported
Asmt Validity	not reported
Asmt Timing	two weeks after (posttest) and 3 months after (delayed post)

SRI International

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

www.sri.com

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

BILL & MELINDA GATES *foundation*