Defining What Empirically Works Best: Dynamic Generation of Meta-Analysis for Computer Science Education

Monica M. McGill
Knox College
Department of Computer Science
Galesburg, IL
mmmcgill@knox.edu

Tom McKlin
The Findings Group
Decatur, GA
tom@thefindingsgroup.com

Errol Kaylor
Knox College
Department of Computer Science
Galesburg, IL
ebkaylor@knox.edu

ABSTRACT

In an effort to evaluate computer science education using more modern, automated data science techniques, we consider Hattie's work in Visible Learning then define a comprehensive framework to provide the capability to automatically generate a quantitative meta-analysis using predefined moderators (e.g., age, grade, etc.) with data derived from multiple individual research studies. To define the initial criteria, we developed a list of critical questions that the framework must address, including what moderators are most important to include, how to address homogeneity across various studies, how to define categories of influencing factors, and how to compute summary effect size. This initial framework describes how the meta-analysis is derived from effect sizes that are calculated based on each mean and standard deviation reported in experimental and quasi-experimental studies. Since the goal of this foundational research is to create an auto-generated meta-analysis tool, we define a basic user experience that would allow users to select moderators and predefined levels of heterogeneity (such as "include only random control group studies" or "include studies reported in journal articles") for inclusion in the meta-analysis. We conducted a feasibility study of the framework using data (number of participants, mean, standard deviation) collected from 21 data samples curated from eight computer science education articles with a primary and secondary focus across ten venues (2012-2018). We consider what we learned conducting the study, including the need for full system transparency, issues related to data integrity, and issues related to defining and selecting appropriate formulas for differing sets of data.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; **Computing education programs**; **Computer science education**.

KEYWORDS

Education, Meta-Analysis, Empirical, Analysis, Effect size, Moderators, Hattie, Visible Learning

ICER '19, August 12–14, 2019, Toronto, ON, Canada

ACM ISBN 978-1-4503-6185-9/19/08…$15.00
https://doi.org/10.1145/3291279.3339401

ACM Reference Format:
Monica M. McGill, Tom McKlin, and Errol Kaylor. 2019. Defining What Empirically Works Best: Dynamic Generation of Meta-Analysis for Computer Science Education. In *International Computing Education Research Conference (ICER '19), August 12–14, 2019, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3291279.3339401

1 INTRODUCTION

[W]here is the knowledge we have lost in information?

T.S. Eliot, "The Rock"

In a recent query of 507 double-blind, peer reviewed articles in primary and secondary computing education, 245 were identified as research articles [30]. Of these, 86 (35.1%) were further identified as quantitative studies reporting data such as mean, standard deviation, t-tests, and the results of other statistical analyses.

In quantitative studies such as these, computing education researchers and evaluators often conduct significance tests that report the presence or absence of a significant difference between two groups or from one time point (pre) to a later time point (post). This is usually reported as a p-value, and educational research often uses the $p<0.05$ benchmark for significance. One common mistake is interpreting p-values to be "more significant" if they are smaller (e.g. that $p<0.001$ is "more significant" than $p<0.05$). However, this value merely detects the presence or absence of a significant difference.

Researchers calculate effect size to measure the magnitude of the difference [11], and dozens of effect size measures exist. They are generally broken down into two categories: those that measure differences in means and those that measure the strength of relations. For measuring mean differences, researchers often use Cohen's *d*, Glass' Δ, or Hedges' g, while measuring the strength of relations often uses R-squared or eta-squared [38].

In 1999, the American Psychological Association (APA) established a task force to investigate banning Null Hypothesis Significance Testing (NHST) from recommended practice, and while the task force did not recommend a ban, they did recommend reporting estimation, such as confidence intervals [24]. The APA's Fourth Edition of the Publication manual of the American Psychological Association (1994) is the first to recommend that researchers report effect size information [2]. The Sixth edition (2010) encourages meta-analyses and provides detailed guidelines for reporting meta-analyses [3]. Taken together, the APA recommends moving away from NHST and toward more cumulative approaches that incorporate effect sizes and meta-analyses.

In fact, Sun, Pan, and Wang (2010) claim that not reporting effect size is detrimental and that effect size should be reported for significant and non-significant findings [38]. Effect size can give us a tool for analyzing the relationship among similar variables across research studies, a practice that uses multiple studies to move the research community closer to a true effect [40]. It also allows for comparisons across findings and also provides data important for conducting secondary analyses such as meta-analyses.

Employing effect size in meta-analyses addresses a critical issue in social science research that has emerged over the past decade: the problem with replicating findings [12, 16]. Several important studies have revealed problems with replicating findings in top journals and the U.S. National Science Foundation recently released guidelines for replicating and reproducing studies in education research [16]. The authors of one such study note that a single study "almost never provides definitive resolution for or against an effect and its explanation" [12, p. 1]; consolidating findings via meta-analysis can show effects over time, different student demographics, and different educational settings to more accurately explain the effects of interventions on student and teacher mediating and outcome variables.

As computing continues to turn to data science and informatics to investigate large sets of data and their relationships, we consider how we can use existing data sets to help prepare computer science education research for the future. Therefore, the overarching research question for this study was:

> What would a framework for automating the aggregation and reporting of the effect of various factors affecting academic achievement in primary and secondary computing education entail?

In other words, is it feasible to create an automated system for dynamically generation meta-analyses in computer science education and if so, what would go into the design of such a system?

This research study is important for establishing early frameworks and criteria for creating data-driven tools for empirically evaluating and comparing results of research studies in computer science education. Through a data-driven approach that takes into account the demographics of various students, needs of schools and districts, duration of the treatment, and/or the intervention itself (like a summer camp for girls) in primary data, we can begin to cultivate a stronger collection of evidence-based practices.

The long-range impact of this study will ultimately affect the students based upon the decisions made by policymakers, curriculum designers, teachers, and program coordinators. This study is also important for those researchers, grant-institutions, and evaluators who recognize the importance of shifting the model from a resource-intensive narrative driven-approach where data from studies are kept on private files stored away from the public eye to one where data is open and accessible so that modern data science approaches can be used to aid in the discovery of how learners learn best.

2 BACKGROUND

Whereas a literature review is primarily a qualitative process with descriptive statistics, a meta-analysis is a statistical analysis of analyses across multiple studies–or as Glass, who coined the term "meta-analysis" in education in 1976, refers to it as an "analysis of analyses" [17]. Creswell states that it is an evaluation of individual results in which an overall numeric index of the magnitude of integrated results is provided [13]. As Glass notes, the integration of the results of hundreds of individual studies can aid with the accumulation of knowledge and evolution of practices. In this section, we provide a brief background in the evolution of meta-analysis, the selection of formulas, Hattie's work in Visible Learning, factors for classifying data, and moderators for defining which studies to include within a meta-analysis. All of these provide a basic and necessary understanding how an automated system might unfold.

2.1 Meta-analysis and Effect Size

Earlier efforts in meta-analysis wholly focused on conducting a qualitative analysis of multiple studies, which have their own limitations. The researcher would read, absorb, and evaluate the studies to be integrated, with different reviewers using different criteria for deciding which studies to include. Borenstein, et al., identify limitations of these types of narrative studies, the first of which is the weighting of studies being performed at the discretion of the reviewer–those critical decision-making processes may or may not be articulated within the article [7]. A second limitation is that these forms of meta-reviews become "less useful as more information becomes available" [7, p. xxii]. As the number of studies grows, a qualitative analysis of these works becomes untenable.

Because of these issues, the process has moved more towards systematic literature reviews and quantitative methods, with well-defined rules and steps for conducting such studies. Studies involving quantitative data include a form of statistical synthesis of the data, where weights are assigned by a mathematical formula. This analysis provides transparency and more objectivity to the results and provides a path for replicating the study.

The guidelines formulating the process are similar to other research steps–beginning with finding and selecting studies, then identifying and coding study characteristics. This is followed with the analysis of the data, and then reporting of the findings.

Several statistical methods have been used to perform this integration of data. Effect size is a standardized difference between two sets of mean and standard deviation, either between an experimental and control group or pre-post results of the same group. In this context, it is used to determine the "strength of the conclusions about group differences." [13] Though there are several ways to measure effect size, one of the common formulas, Cohen's *d*, can measure the effect between two groups using mean, standard deviation, and number of participants. In the field of psychology, Cohen has determined 0.2 as a small effect size benchmark, 0.5 a medium effect, and 0.8 a large effect [10].

Since a single study only provides one point of reference, one study can rarely illustrate the true effects of an intervention on students. Even with replication, however, reviewing studies individually can be problematic. For example, in a 2015 study, 270 psychologists sought to replicate 100 studies in top journals and found that only about 40% of those studies could be replicated [12]. Further, the successful replications showed weaker effect sizes than the original. In another study, researchers tried replicating 21 highly-influential social science studies and found that only 13

could be replicated [8]. The replication attempts had sample sizes five times greater than the original and had effect sizes that were about half the size of the effect sizes in the original studies. The integration of data, therefore, may provide a stronger indication of how effective treatments and interventions are.

We also note that when planning a study for potential acceptance into the What Works Clearinghouse (WWC), the WWC Standards Handbook, Version 4.0 recommends that researchers calculate the Minimum Detectable Effect Size (MDES), the smallest true effect, in standard deviations, of the outcome for a given level of power and statistical significance [33]. They set 0.25 as the minimum threshold for substantively important effects and that projects need to be able to detect effects at or below 0.25. This means that researchers, at the outset, should design studies with a minimum detectable effect size no greater than 0.25.

2.2 Hattie's Visible Learning Model

Hattie has introduced and maintained a massive secondary data analysis project involving the review of over 1,200 meta-analyses, primarily for primary and secondary education but also with implications for higher education [19, 20, 29]. Through this work, Hattie and others approximate through research that the

> ...effect size of 0.40 (calculated with Cohen's *d*) indicates that students have gained at least a year's worth of growth for a year in school. The implication is that 0.40 should be the expectation for instruction and intervention. An effect size lower than 0.40 suggest that the instruction or intervention was less than effective and may warrant changes or revision. At the very minimum, an effect size below 0.40 begs for discussion about the effort. [1, p. 162]

Thus, an effect size of 0.4 becomes the benchmark and differentiates between expected growth due to maturity versus actual impact from a curriculum, program, or activity. This benchmark and the analysis techniques extend beyond general primary and secondary education into the differentiated fields including reading, STEM, science, and mathematics [1, 19, 21, 22].

Though there is fair criticism to Hattie's work [6, 37], part of our interest in it centers on his efforts of applying statistical measures to integrate large bodies of secondary data, the potential for benchmarking effect size in meta-analysis studies, the categorization of influencing factors, and the moderators that help define heterogeneity across studies. It is also worth noting that Hattie's work is a *meta-analysis of hundreds of meta-analyses*, while our effort is to move towards the first step of generating *a single meta-analysis*.

2.3 Factors

Hattie's work includes the identification of domains and subdomains in order to make comparisons across the spectrum of the 1,200 meta-analyses that he has reviewed. Hattie identifies six domains: Teacher, Student, Curricula, Classroom, Teaching Approaches and Home [19]. Each of these are further dissected into unique subdomains. Though all subdomains and the particular influencers are too long to list, an example for the Student domain includes subdomains *Background*, *Attitudes* and *Dispositions*, *Physical influences*, and *Preschool experiences*. The subdomain *Background* includes *Prior achievement*, *Pieagetian programs*, *Self-report grades*, and *Creativity*. These contributions to academic achievement from the student have all been studied in meta-analyses and Hattie has taken that work to integrate it and analyze it across the data in the body of meta-analyses.

His dataset takes a broad approach to all data reported in meta-analysis. This classification of influencers of academic achievement in students is not unique and other formal models exist. Farrington, et al., identify five categories of noncognitive factors (influencers), including academic behaviors, academic perseverance, academic mindsets, learning strategies, and social skills [14]. Marzano's model was developed based on identifying the most significant effect sizes across multiple studies. Three major groupings were identified– student-level, teacher-level, and school-level, with each level further broken down into influencing factors [28].

Looking at smaller sets of data that are typically reported in computer science education research, we turn to the Lee and Shute model as a classification system [25, 26]. The Lee and Shute model was used for previous analysis of constructs measured across evaluation instruments for its well-defined, rich, overall structure for classification of noncognitive factors in learning [31]. For brevity, the individual factors are not listed here, but the major components and subcomponents are Student-Personal (Student Engagement, Learning Strategies) and Social-Contextual (School Climate, Social-Familial Influences).

2.4 Moderator Variables

Marzano (1998) provides eight moderator variables, or those variables that "influence[s] the strength of a relationship between two other variables" [4, 27, p. 1774]. These include whether the technique was designed for the teacher or student, the degree of specificity of the influence, grade level of students, student ability, treatment duration, specificity of dependent measures in the treatment, methodological quality, and publication type.

In this context, researchers can use such moderators and other independent variables to help define the level of heterogeneity across studies. In a meta-analysis, one could consider all studies where students were in high school and the treatment of duration was one semester. Given that Marzano includes publication type as a moderator, we can also identify other variables related to the publication and the research study to further create more heterogeneity when choosing studies to include in a meta-analysis.

2.5 Summary

Glass makes the case that how we define effective practices in educational research may not be the most appropriate [17]. Sometimes those practices are defined through vote-taking, in narrative literature reviews or by reading conflicting results of studies. Sometimes they are defined by the researcher's reputation. And sometimes they are defined by those with the largest voices, the largest followings, and the largest reach.

Having meta-analysis (using empirical data) readily available in computer science education is important if we want to consider the full integration of studies and how these studies either complement or contradict each other and why. The process of creating tools

to conduct this meta-analysis must be considered thoughtfully, pragmatically, and transparently.

Although Hattie's work is a meta-analysis of over 1,200 meta-analyses, in the field of computing education we do not have sufficient meta-analyses studies to replicate the Visible Learning model. However, we currently have a system for storing and evaluating data reported in articles. Given this, what if we created a tool to automatically generate meta-analysis based on sets of data whose heterogeneity is set by the user? This research explores this possibility, with the next two sections describing a framework for such a tool and a manually-generated feasibility test of the framework.

3 DEVELOPMENT OF FRAMEWORK

Based on the above research, we developed the following questions each of which must be addressed and built into the framework:

- What moderators need to be defined and extracted from the articles?
- What is the intervention?
- What defines a) homogeneity and b) category of "factors"?
- To compute the summary effect size, a) what values do we need and b) what formula do we use? Are there other issues related to effect size (e.g., minimum detectable effect size) that should be considered?
- For an automated system, what would the user interface look like to be most useful for users? What are effective ways to present the results?

After investigating several models for performing a meta-analysis, we chose to use Basu's steps on conducting a meta-analysis as the methodology that informs our framework [5]. Basu recommends the following nine steps for conducting a meta-analysis:

(1) Frame a question (based on a theory)
(2) Run a search (on Pubmed/Medline, Google Scholar, other sources)
(3) Read the abstract and title of the individual papers.
(4) Abstract information from the selected set of final articles.
(5) Determine the quality of the information in these articles. This is done using a judgment of their internal validity but also using the GRADE criteria
(6) Determine the extent to which these articles are heterogeneous
(7) Estimate the summary effect size in the form of Odds Ratio and using both fixed and random effects models and construct a forest plot
(8) Determine the extent to which these articles have publication bias and run a funnel plot
(9) Conduct subgroup analyses and meta regression to test if there are subsets of research that capture the summary effects

Several of these steps will be integrated into the system and automated, including steps 2, 4, and 7. To provide a natural division between the non-automated and automated features, in the next section we present the process from the user perspective first, followed by the technical (automation) requirements for the system.

3.1 User Perspective

After considering the questions and steps above, we folded them together and extracted the technical details (discussed in the technical specifications in section 3.2). We propose the following steps strictly from the user perspective.

3.1.1 Develop a research question for which the meta-analysis will be conducted. As a start, the user should frame the research question for which the meta-analysis will be conducted. In order to answer the user's research question using the automated meta-analysis generator, the research question should be framed in the context of available factors presented in the system.

3.1.2 Select the factor for which the meta-analysis will be generated. The factors will be presented as cognitive factors (content knowledge–e.g. computational thinking, robotics, programming concepts, etc.) and noncognitive factors. Noncognitive factors will be framed in the context of the Lee & Shute Model (e.g., self-efficacy, interest/curiosity, grit) [25]. Only factors with data available in the system will be presented for selection. The user will be required to select one of the factors.

3.1.3 Choose the level of heterogeneity based on article-related data. The user will be presented with all articles that measure the selected factor. The user can include all articles or select those within a specified range of publication dates (i.e., all articles published within the last three years), publication venues (i.e., only journal articles), or other relevant article data. The quality-control of the data will be driven by the user based on the user's own research, desired level of integrity, type of experiment, and comfort with the data from the particular venues.

Provided with this selection process will be links to the library card for each article to determine if the article meets the user's needs. The library card will include all relevant information manually curated from the article, including intervention and the number of participants in the study.

3.1.4 Choose the level of heterogeneity based on moderator specific data. The user will be presented with available moderators for the selected factor and will choose which moderators will be included in the meta-analysis. The moderators will include items such as grade level of students and whether the intervention was designed for students or teachers.

3.1.5 Conduct the meta-analysis. Once the levels of heterogeneity are set by the user, the user will then have the option of running the meta-analysis.

3.1.6 Review the results of the meta-analysis. The results of the meta-analysis will include a forest plot and the summary effect size presented graphically. The default method of analysis will be the random-effects model; however, the user will be able to change this to a fixed-effects model. Each individual intervention included in the meta-analysis will be listed with a link to the article's library card, the specific intervention, and the individual effect size. This data will be ranked in order of the weight (highest to lowest) used in the calculation.

In addition to this information, the user will be presented with definitions and analysis assumptions used in the calculations.

3.1.7 Refine the levels of heterogeneity in source data (articles) and moderators. After the results are presented, the user will be able to modify the heterogeneity of the articles and the moderators to further refine the effect size.

3.2 Technical Specifications

This section provides an overview of the technical specifications needed to support the end-user functionality as well as meet standards for educational research.

3.2.1 Identification of articles to include in the data set. The data set will come from all the manually curated and vetted data that is part of the csedresearch.org dataset [30]. An explanation of the full methodology for identifying articles can be found in []. Briefly, the csedresearch.org dataset currently contains data from over 500 primary and secondary computing education articles published from 2012 to 2018 across ten venues. All of these venues utilize a double-blind, peer-review process.

Although statistical data is not currently included in the database, the dataset will be extended to include the intervention, specific statistical data (number of participants, mean, and standard deviation), identification of each moderator for the intervention, and the cognitive (content knowledge) and noncognitive (e.g., self-efficacy, interest/curiosity, grit) factors being measured.

3.2.2 Identification of moderators to be included. As a model, the type of report produced will be dependent on the variables selected by the user. These variables would produce homogeneity at one basic level. Therefore, if the user selected secondary education (e.g., U.S. grades 9-12), the effect sizes (and informing variables) would be displayed only for those grades.

Given the wealth of data already curated from the articles, it is conceivable that many independent variables can operate as a moderator–for example, prior experience of students, gender, student socioeconomic status, and students with disabilities.

3.2.3 Identification of formulas to be used. To identify the summary effect size formulae to use, we consider both fixed and random effect measures. Borenstein, et al., argue that the fixed effect model should be used if all studies in the analysis are functionally identical and the goal is to compute the common effect size for the identified population only [7]. The random effects model is best when accumulating data from a series of independent studies from various researchers and when wanting to generalize to a range of different scenarios.

In addition, individual effect sizes for each construct/item will be calculated uniformly across the meta-analysis. We need to identify the most accurate formula for calculating these effect sizes. For example, for pretest-posttest-control group designs, previous research shows that an "effect size based on the mean pre-post change in the treatment group minus the mean pre-post change in the control group, divided by the pooled pretest standard deviation" is a preferred model [32]. Pooled standard deviation may also be strategic to use as the denominator for calculating effect size. Each method of calculation will need to be thoroughly evaluated and the formulas will need to be easily viewed by the user [7, 18].

4 METHODOLOGY

To test the feasibility of the above framework, examine its nuances, and redefine its steps, we conducted a feasibility study using the primary and secondary computer science education data from the csedresearch.org dataset. This section describes our methodology as well as the results of the study.

4.1 Technical Specifications

Beginning with the 507 articles, we first manually examined each article individually to determine what statistical measures were reported. Once these measures were recorded in the database, we queried the database to find only those articles that were tagged as reporting both mean and standard deviation. This resulted in 75 articles.

We then reexamined each of the 75 articles and eliminated articles not meeting the inclusion criteria:

- Reports data from within-groups or control-experimental studies, Clearly stating mean and standard deviation for data relevant to the the intervention/treatment,
- Reports sample sizes for this data,
- Reports effects of an intervention on students, as opposed to teachers or others.

Articles in which means and standard deviations were only reported for sample composition (gender, race, year in school) were excluded, along with articles that showed this data only graphically. Four were eliminated due to improper coding in the database. In others, there were no within-groups or control/experimental data compared. Eight articles did not report the number of participants, means, and/or standard deviations for each group, while two compared results via attributes like gender or race. After eliminating articles that did not fit this criteria, we looked at our remaining 20 studies and classified all 103 factors measured in each set of data according to the Lee and Shute model [25].

We then went through the refinement process for the feasibility study. We chose to run our pilot on the most-reported factor for the most reported type of study in order to have the most robust set of data. The most reported type of study was the One-Group Pre-test Post-test group. We then assigned the factors to each and determined that Interest/Curiosity was the most examined factor. This resulted in a final set of eight studies with 21 sets of data (mean, standard deviation, number of participants). We then added the data for each of the moderators specified by Marzano and, as we progressed in our analysis, noticed a need for an additional two moderators (see Table 1).

Within our data, we found that a single article may describe and analyze several different interventions, and we view these interventions as unique and separate. The results were not consolidated into one effect size, even if the factor being measured was the same. Likewise, one intervention may result in an analysis of multiple survey questions all investigating the same factor. However, we did not group these factors together into one construct. This is an important consideration and is discussed further in the Discussion section below.

4.2 User specifications

We followed the framework in section 3.1. Each step is defined here.

Table 1: Marzano's and our Custom Moderators (* denotes data already curated from articles

Moderators defined by Marzano
Technique/Intervention Designed for Student or Teacher
Specificity of Influence
Grade Level of Students*
Student Ability
Duration of Treatment
Specificity of Dependent Measures
Methodological Quality*
Publication Type*
Custom Moderators
N-values Reported (Clean/Not Clean)
Evaluation on Student or Teacher

4.2.1 Develop a research question for which the meta-analysis will be conducted. For this feasibility study, we chose to select a factor that had the most robust set of data for testing the framework. Since in the case of this particular study we were very familiar with the data, it, in turn, drove the research question for this feasibility study: *What are the most impactful interventions for increasing interest among middle and high school students in computing?*

4.2.2 Select the factor for which the meta-analysis will be generated. As noted above, the factor we selected to investigate is Interest/Curiosity which falls under the Student Engagement component and the Emotional Engagement (Affect) subcomponent according to the Lee and Shute model.

4.2.3 Choose the level of heterogeneity based on article-related data. For this feasibility study, we chose to only include studies that had experiment types of "One-Group Pretest-Posttest design" and with a factor of Interest/Curiosity, since this turned out to be the most popular type of study in our set of data and thus be able to yield the most robust set of data for this analysis.

4.2.4 Choose the level of heterogeneity based on moderator specific data. Since this study was for those interventions affecting and evaluating students, we removed all studies that had a teacher, preservice teacher, or mentor focus. As it turned out, for the articles in this group, the studies only investigated middle school and high school students, which would be considered a moderator as well.

4.2.5 Run the meta-analysis. For this feasibility study, we conducted the data analysis from the 21 sets of data using the Review Manager 5.3 (RevMan) tool [9].

4.2.6 Review the results of the meta-analysis. After entering the data into RevMan, we were presented with the individual effects, the summary effect, confidence interval, and a forest plot. Since we calculated the results using the Random Effects model, we also chose to view the forest plot by weighted effects.

4.2.7 Refine the levels of heterogeneity in source data (articles) and moderators. No changes were initially made, though we discuss this in further detail in the Discussions section.

4.3 Results

In Figures 1 and 2, we present the results in a form the the user could see. To make the formulas transparent to the user, all assumptions made will be explained as well as the specific formulas used. Forest plots are an important way to visually present the results of the comparative analysis and will be included, when appropriate. The Summary Effect Size graphic could be presented in terms of the criteria established through Hattie in the Visible Learning project. At a minimum, weak, moderate, and strong effect size values could be provided for contextualization and benchmarking.

In addition, below the Summary Effect graphic, each intervention that was part of this analysis will be displayed along with a link to the article summary, ranked in order of weight.

5 DISCUSSION

The steps in the feasibility study were easy to follow. Although the question and selection of moderators were driven by our knowledge of the resulting data for the purpose of defining the most robust set of data we could for this study, we postulate that for a researcher or evaluator this process would be ultimately straightforward. As we examine the technical specifications, including the data and the analytics to be implemented into such a system, we consider the data analytics, the data integrity, and what "full system transparency" may entail so a user can meaningfully interpret the results.

5.1 Data Analysis

The default equation for the meta-analysis will be the random effects model. However, there may be times when a fixed effects equation is more appropriate. For this reason, these models will need be to easy to select. Sufficient information about the models should be provided so the user can understand their differences.

In our feasibility study, if a study reported n, mean, and standard deviation for each question in a survey that related to interest/curiosity, we treated each as individual entries to be analyzed. Another choice we could make is to automatically condense these results into one construct. In the former, there is a bias towards these studies since they are over represented for each factor. After further consideration, grouping these into one construct would be a more preferred path. It may be possible to even calculate the reliability of the group by calculating Cronbach's alpha–even going so far as to automatically including only data from those questions that generate a Cronbach's alpha greater than 0.70.

We also note here that Hattie's 0.4 benchmark is primarily for student learning over the course of one year. Further investigation is needed to see if this might be a useful benchmark for interventions that are significantly shorter in duration. We also must consider what type of benchmarks could be used for interventions designed to change participants' perceptions or attitudes, such as interest in computing–or if we leave these interpretations completely open for the user.

On another interesting note, moderators could serve as a way to select subgroups within a meta-analysis. In fact, as we move to a data-mining model, analyses can be auto-generated based on known moderators stored in the data without any human-driven selections. These analyses could run quietly in the background,

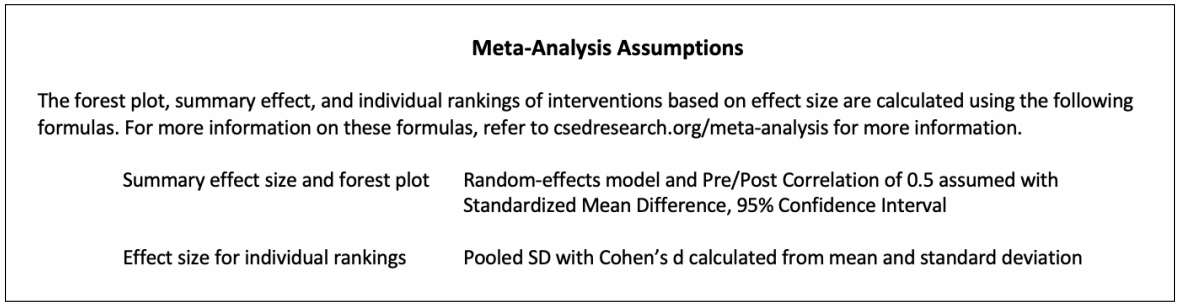


Figure 1: Assumptions provided to the user.

notifying users through email, an app, or even social media of interventions that may be more impactful for given subgroups.

5.2 Data Integrity

The data becomes a central, ongoing point of concern for this framework. There are many experimental and quasi-experimental studies in the database already and more will be added in the coming months and years; however, only data from eight studies qualified for this feasibility study. We reported above that over 10% of the articles did not report sufficient data to compare the pre-post results or the experimental-control group results. Some reported means and number of participants without standard deviation. Some reported means and standard deviations without explicitly stating the number of participants. Some reported "no significant difference found" without reporting the actual values (Mean, SD, or N)–when in fact, the data can play a significant role in the larger context of a meta-analysis. The "no significant difference found" data cannot be represented in any meta-analysis, since there is no data provided for inclusion. Further, with few appropriate sets of data for comparison, susceptibility to bias across the studies creeps in. There is, therefore, a need to either engage researchers in reporting more precise data in their reports or shifting the model to engage researchers in providing raw data directly to a central database system, like the Open Science Framework [15].

Further, more research should be conducted into the minimum detectable effect size as well as the fail-safe N in meta-analysis for assessing publication bias [33, 34]. The integration of these and additional data cleaning practices in data analysis and mining can further ensure that clean, acceptable data is included in the meta-analysis [35]. This can play a part in reducing bias and could be integrated into the technical specifications before big data arrives.

5.3 System Transparency

Finally, and potentially foremost, the system should be fully transparent to enable the user to make wise decisions about the Summary Effect Size and other data points that are generated from the meta-analysis. Although any data can be statistically misrepresented, the fault of any misrepresentation should not necessarily fall on the system. By providing full transparency, the user can be more assured that they understand what the values mean–and don't mean. Any such system should also consider if and how limitations of the resulting meta-analysis can be presented. This must be one of the primary pillars of the system's technical and user specifications.

In considering the above, there is a potential to have the results of the meta-analysis and any known limitations presented in narrative form. The move from purely data into natural language generation for human-readers has already begun and is being used commercially across some fields–for example, sports reporting and insurance companies [23, 36] through the use of a tool such as Tableau [39]. These same techniques could be applied in a meta-analytics system for computer science educational research.

6 CONCLUSION

Glass ends his well-read 1976 article that introduces meta-analysis with this: "Extracting knowledge from accumulated studies is a complex and important methodological problem to which I commend your attention" [17, p. 8]. This research is a first-step in comparing data presented in individual computer science education research studies in a systematic and integrated way through an automatically generated meta-analysis.

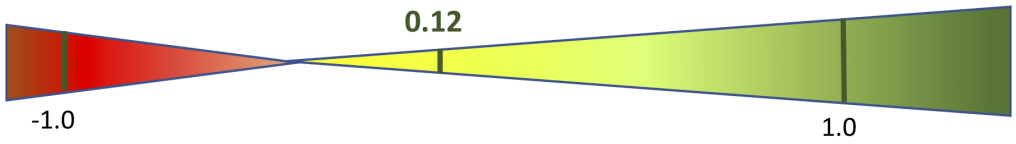
The motivation behind this work is to help prepare computer science education research for the future. As studies and data across the educational community become more abundant with richer data sets, having a tool in place that will automate the integration of data from individual studies becomes important. Having a system in place that takes this one step farther–quietly runs such calculations as we sleep in order to find best practices for particular groups of learners–is the future in many fields already. There is every reason to believe that preparing our data and finding and testing appropriate calculations will make the transition of human-generated data to fully automated systems much smoother. The ability to test such work now, before the data being generated becomes too voluminous, unwieldy, disorganized and chaotic for anyone to control, positions the computer science educational research community much better for the future.

We recognize that Hattie's work on Visible Learning is a multi-year effort that evaluates meta-analytic studies. Our framework is a first-step in producing the type of meta-analytic studies that can be compared and contrasted for various demographic groups for decades to come [19, 21, 22].

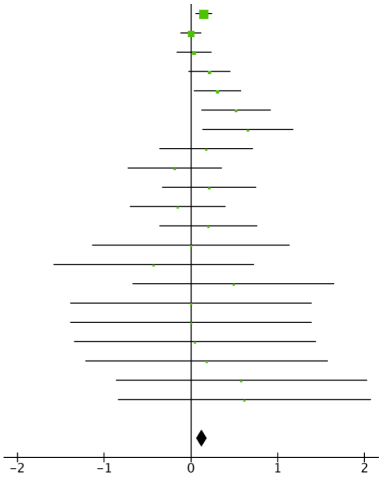
ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 1757402. csedresearch.org is further supported by U.S. National Science Foundation under Grant Nos. 1625005, 1625335, 1757402, and 1745199.

Summary Effect Size for Interest/Curiosity among Middle and High School Students



Forest Plot for Student Engagement -> Emotional Engagement (Affect) -> Interest/Curiosity Middle and High School Students



Pre-test N = 2,303
Post-test N = 2,213

Summary Effect Size with Confidence Interval
0.12 [0.06, 0.18]

Heterogeneity
$Tau^2 = 0.00$
$Chi^2 = 19.48$, df = 20 (P = 0.49)
$I^2 = 0\%$

Test for overall effect
Z = 3.97 (P < 0.0001)

Individual Rankings of Interventions based on Weight (For Random-Effects)

Intervention	Effect Size	Weight	Author(s)
One-week summer camp to teach programming concepts	0.145	42.3%	Guzdial, Ericson, McKlin, Engelman (2014)
One-week summer camp to teach programming concepts	0.000	25.8%	Guzdial, Ericson, McKlin, Engelman (2014)
STEM Summer course, including an optional AP CS A prep course	0.037	8.9%	Scott, Martin, McAlear, Koshy (2017)
Game Programming	0.522	6.2%	Lakanen, Isomöttönen, Lappalainen (2012)
Trained near-peer mentors for a camp to teach how to program using App Inventor	0.306	4.7%	Clarke-Midura, Poole, Pantic, Hamilton, Sun, Allan (2018)
Game Programming	0.213	2.2%	Lakanen, Isomöttönen, Lappalainen (2012)
Summer camp with GET IT presentation	0.664	1.3%	Outlay, Platt, Conroy (2017)
Computer Science Camp (for girls)	0.177	1.2%	Won Hur, Andrzejewski, Marghitu (2017)

< Click here to see more interventions >

Figure 2: Representation of a summary page with statistics and graphs.
The analysis assumptions shown in Figure 1 will appear above this.

REFERENCES

[1] John Almarode, Douglas Fisher, Nancy Frey, and John Hattie. 2018. *Visible Learning for Science: What Works Best to Optimize Student Learning.* Corwin.

[2] American Psychological Association. 1994. *Publication manual of the American Psychological Association (4th ed.).* American Psychological Association: Washington, DC, USA.

[3] American Psychological Association. 2010. *Publication manual of the American Psychological Association (6th ed.).* American Psychological Association: Washington, DC, USA.

[4] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

[5] Arindam Basu. 2017. How to conduct meta-analysis: a basic tutorial. (2017). https://peerj.com/preprints/2978v1.pdf

[6] Pierre-Jérôme Bergeron and Lysanne Rivard. 2017. How to engage in pseudoscience with real data: A criticism of John HattieâĂŹs arguments in visible learning from the perspective of a statistician. *McGill Journal of Education/Revue des sciences de l'éducation de McGill* 52, 1 (2017), 237–246.

[7] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2009. *Introduction to meta-analysis.* Wiley Online Library.

[8] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637.

[9] TC Cochrane. 2008. Review Manager (RevMan) 5.3. *Copenhagen: The Nordic Cochrane Centre* (2008).

[10] Jacob Cohen. 1977. Statistical power analysis for the behavioral sciences, Rev. (1977).

[11] Jacob Cohen. 1990. Things I have learned (so far). *American psychologist* 45, 12 (1990), 1304.

[12] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[13] John W Creswell. 2008. Qualitative, quantitative, and mixed methods approaches.

[14] Camille A Farrington, Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W Johnson, and Nicole O Beechum. 2012. *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance–A Critical Literature Review.* ERIC.

[15] Erin D Foster and Ariel Deardorff. 2017. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 105, 2 (2017), 203.

[16] National Science Foundation. 2018. Companion Guidelines on Replication & Reproducibility in Education Research. https://nsf.gov/pubs/2019/nsf19022/nsf19022.pdf

[17] Gene V Glass. 1976. Primary, secondary, and meta-analysis of research. *Educational researcher* 5, 10 (1976), 3–8.

[18] Jessica Gurevitch. 1993. Meta-analysis: combining the results of independent experiments. *Design and analysis of ecological experiments* (1993).

[19] John Hattie. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* routledge.

[20] John Hattie. 2015. The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology* 1, 1 (2015), 79.

[21] John Hattie, Douglas Fisher, Nancy Frey, Linda M Gojak, Sara Delano Moore, and William Mellman. 2016. *Visible learning for mathematics, grades K-12: What works best to optimize student learning.* Corwin Press.

[22] John Hattie and Gregory CR Yates. 2013. *Visible learning and the science of how we learn.* Routledge.

[23] Automated Insights. 2018. Allstate presents Wordsmith Extension at 2018 Tableau Conference. https://automatedinsights.com/blog/how-allstate-uses-natural-language-generation-directly-in-tableau-dashboards/

[24] David H Krantz. 1999. The null hypothesis testing controversy in psychology. *J. Amer. Statist. Assoc.* 94, 448 (1999), 1372–1381.

[25] Jihyun Lee and Valerie J Shute. 2010. Personal and social-contextual factors in K–12 academic performance: An integrative perspective on student learning. *Educational Psychologist* 45, 3 (2010), 185–202.

[26] Jihyun Lee and Lazar Stankov. 2018. Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences* 65 (2018), 50–64.

[27] Robert J Marzano. 1998. A theory-based meta-analysis of research on instruction. (1998).

[28] Robert J Marzano. 2003. *What works in schools: Translating research into action.* ASCD.

[29] Deb Masters, Kate Birch, and John Hattie. 2015. *Visible learning into action: International case studies of impact.* Routledge.

[30] Monica M. McGill and Adrienne Decker. 2017. Computer Science Education Repository. https://csedresearch.org

[31] Monica M McGill, Adrienne Decker, Tom McKlin, and Kathy Haynie. 2019. A Gap Analysis of Noncognitive Constructs in Evaluation Instruments Designed for Computing Education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education.* ACM, 706–712.

[32] Scott B Morris. 2008. Estimating effect sizes from pretest-posttest-control group designs. *Organizational research methods* 11, 2 (2008), 364–386.

[33] US Department of Education. 2008. WWC Procedures and Standards Handbook: Version 4.0.

[34] Robert G Orwin. 1983. A fail-safe N for effect size in meta-analysis. *Journal of educational statistics* 8, 2 (1983), 157–159.

[35] Jason W Osborne and Amy Overbay. 2013. Best practices in data cleaning. *Best practices in quantitative methods* (2013), 205–213.

[36] Associated Press. 2016. Press Release: AP expands Minor League Baseball coverage. https://www.ap.org/press-releases/2016/ap-expands-minor-league-baseball-coverage

[37] Ivan Snook, John O'Neill, John Clark, Anne-Maree O'Neill, Roger Openshaw, et al. 2009. Invisible learnings?: a commentary on John Hattie's book-'Visible learning: a synthesis of Over 800 meta-analyses relating to achievement'. *New Zealand journal of educational studies* 44, 1 (2009), 93.

[38] Shuyan Sun, Wei Pan, and Lihshing Leigh Wang. 2010. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology* 102, 4 (2010), 989.

[39] Tableau. 2019. Tableau. https://www.tableau.com/

[40] Bruce Thompson and J.W. Osborne. 2008. Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. *Best practices in quantitative methods* (2008), 246–262.