MIND THE GAP: ASSESSING CURRENT PEDAGOGICAL APPROACHES TO TEACHING EFFECT SIZE ESTIMATION AND INTERPRETATION IN UNIVERSITY LEVEL PSYCHOLOGY STATISTICS CLASSROOMS

MARK CHRISTOPHER ADKINS

A DISSERTATION SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

SEPTEMBER 2026

© MARK CHRISTOPHER ADKINS

ABSTRACT

Effect size estimation and interpretation are fundamental to statistical reasoning and evidence-based research, yet their coverage in university statistics instruction is not well understood. This dissertation examined instructional practices among Canadian university social science statistics instructors and developed an online educational resource to support teaching these concepts.

Study 1 surveyed instructors about their inclusion of eight effect size techniques in undergraduate and graduate statistics courses. For all excluded techniques, instructors were asked to list the barriers which led to the exclusion of each specific technique within their classrooms. These findings informed the development of an interactive website designed to address the identified inclusion barriers/challenges and consequently make it easier to integrate the interactive activities into existing curricula.

In Study 2, instructors evaluated the interactive website across seven instructional design dimensions, including navigation, graphical representation, organization, content utility, purpose clarity, simplicity, and readability. The website received positive evaluations across all dimensions, with qualitative feedback informing recommendations for future improvements.

Together, these studies provide insight into current instructional practices and demonstrate the value of a theory-informed educational resource for improving the accessibility, consistency, and quality of effect size instruction in Canadian university statistics education.

DEDICATION

I dedicate this to my amazing wife, Sara, who is still unquestionably the funniest person I know. She is by far the most patient and empathic person I have ever met, and without her unwavering love and support this present work would not have been possible. I also dedicate this to my “Little Man”. You are a fantastic person, and the best code tester a father could ask for. I cannot begin to express the gratitude that I have for you both.

TABLE OF CONTENTS

ABSTRACT ii
TABLE OF CONTENTS iv
LIST OF TABLES vii
LIST OF FIGURES viii
CHAPTER ONE: GENERAL BACKGROUND AND THEORETICAL CONTEXT 1
Effect Sizes in Psychological Research 1
A Gap Between Reporting Standards and Statistical Education 3
Student Difficulties and Misconceptions About Effect Sizes 5
Instructional Challenges in Teaching Effect Size Interpretation 6
Technology-Supported Learning as a Potential Solution 6
Overview of the Current Two-Study Research Program 7
CHAPTER TWO: ASSESSING THE PEDAGOGICAL LANDSCAPE 11
Purpose and Research Questions 11
Preregistration Plan 11
Methods 13
Participants 13
Study 1 Survey Instrument 13
Data Collection Procedures 15
Results 16
Recency of Statistical Training 16
Current Academic Position 17
Level of Instruction 18
Comfort Teaching Effect Size Interpretation and Estimation 19
Perceived Importance of Effect Sizes 20
Effect Size Understanding Among Instructors 20
Patterns of Instructional Coverage in Undergraduate Courses 22
Barriers to Teaching Effect Size Techniques in Undergraduate Instruction 24
Patterns of Instructional Coverage in Graduate Courses 24

Barriers to Teaching Effect Size Techniques in Graduate Instruction 26
Discussion 27
Instructor Values and Conceptual Readiness 27
Instructional Coverage Relative to Preregistered Expectations 27
Barriers as Structural Rather Than Conceptual 28
Scope of Instructional Impact 29
Implications for Study 2 and Instructional Design 30
CHAPTER THREE: HISTORICAL AND THEORETICAL FRAMING 32
Introduction 32
Evolution of Interactive Online Resources in Statistics Education 33
From Computational Instruction to Statistical Reasoning 33
Emergence of Technology-Supported Statistics Learning 34
Interactive Learning as a Response to Persistent Statistical Misconceptions 35
Contemporary Principles for Interactive Educational Resources 36
Implications for Effect Size Education 37
Statistics Education Framework 38
Statistical Reasoning as the Primary Learning Goal 38
Data-Based Learning Through Interaction 38
Instructional Design Framework 39
Constructivist Learning Principles 39
Active Learning and Immediate Feedback 40
Scaffolding Statistical Interpretation 40
Translating Study 1 Findings Into Design Requirements 47
CHAPTER FOUR: ASSESSING ONLINE INTERACTIVITIES 51
Introduction 51
Background 51
Technologies 52
Methods 55
Participants 55

Recruitment 55
Sample Size 55
Current Academic Position 56
Years Since Degree 56
Teaching Rank 57
Analysis 57
Results 59
Website Endorsement 59
Website-Level Design Features 59
Content Utility Across Techniques 60
Purpose Clarity 61
Readability 61
Simplicity 62
Summary of Design Evaluations 62
Discussion 63
Limitations 65
CHAPTER FIVE: GENERAL DISCUSSION, LIMITATIONS, AND FUTURE DIRECTIONS 67
General Discussion 67
Limitations 70
Future Research 71
Concluding Remarks 72
REFERENCES 73
APPENDIX A: STUDY 1 QUESTIONNAIRE 83
APPENDIX B: STUDY 2 QUESTIONNAIRE 92

LIST OF TABLES

Table 1 Summary of Effect Size Techniques

Table 2 Years Since Most Recent Degree

Table 3 Current Academic Position

Table 4 Comfort Teaching Effect Size Interpretation and Estimation

Table 5 Item level accuracy relating to effect size misconceptions

Table 6 Undergraduate-Level Predicted Technique Prevalence Rates and Observed Prevalence Rates

Table 7 Graduate-Level Predicted Technique Prevalence Rates and Observed Prevalence Rates

Table 8 Translation of Study 1 Findings and Literature Themes to Website Design Features

LIST OF FIGURES

Figure 1 Screenshot of Simulation-Based Interactive Activity
Figure 2 Updated Screenshot of Simulation-Based Interactive Activity
Figure 3 Screenshot of Simulation-Based Interactive Activity for Long-Run Probabilities
Figure 4 Screenshot of The Full Monty Hall Simulation
Figure 5 Screenshot of the Long-Run Probabilities for Different Door Selection Strategies
Figure 6 Screenshot of the Build-A-Simulation Activity
Figure 7 Screenshot of a Customized Monte Hall Simulation Activity

CHAPTER ONE:
GENERAL BACKGROUND AND THEORETICAL CONTEXT

Effect Sizes in Psychological Research

Geoff Cumming (2014) has long used the term *the new statistics* in reference to "recommended practices, including estimation based on effect sizes, confidence intervals, and meta-analysis" (p. 7). Central to this framework is the emphasis on quantifying effects and shifting focus away from statistical significance testing. Although definitions of effect size vary, effect sizes have been described as "the amount of anything of interest" (Cumming & Fidler, 2009), as "some magnitude (or size) of the impact (or effect) of a predictor on an outcome variable" (Pek & Flora, 2018), and as a "quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley & Preacher, 2012). Across definitions, two core elements recur: magnitude and relevance to a research question.

Effect sizes therefore quantify the strength of a phenomenon in a way that is directly interpretable in relation to substantive research questions. They are often associated with the notion of *practical significance*, which refers to whether an observed effect is meaningful in real-world or theoretical terms when interpreted within the context of a study. In contrast to null hypothesis significance testing (NHST), which evaluates whether an effect is statistically distinguishable from zero under a specified model, effect sizes address the question of how large an effect is. While discussions of the conceptual limitations of NHST are well documented elsewhere (e.g., Gigerenzer, 2004; Krueger, 2001), such critiques are beyond the scope of this present work. What is central here is that effect sizes possess intrinsic value independent of any single

inferential framework. They can be computed and interpreted alongside, or independent of, significance testing.

Although the rise in prominence of effect sizes in recent decades has sometimes been framed as a response to concerns about NHST (Huberty, 2002; Wilkinson et al., 1999), their history extends back to at least the mid-twentieth century (Huberty, 2002). Effect sizes are indispensable because they provide quantitative answers to scientific questions about magnitude—answers that are essential for theory development, comparison across studies, and cumulative knowledge building.

One clear illustration of their utility appears within the context of statistical power analysis. Effect sizes are a required component of any power calculation, as power depends jointly on sample size, significance level, and the magnitude of the effect under investigation. Power analyses are commonly recommended or required by funding agencies, professional organizations, and journal editors as a means of informing sample size planning and reducing inefficient allocation of research resources (Lakens, 2022). Cohen (1962) demonstrated that many studies in the behavioral sciences were substantially underpowered, leaving researchers with “a relatively (or even absolutely) poor chance of rejecting their major null hypotheses” (p. 151). Subsequent decades have continued to document challenges related to underpowered designs and the selection of effect size estimates using insufficient justification (Cohen, 1992; Albers & Lakens, 2018; Beribisky et al., 2025). These issues underscore the central role of effect size estimation not only in interpreting results, but also in designing rigorous and efficient research.

Although reporting effect sizes has been recommended in psychological and social science research for decades, empirical reviews suggest that it is still inconsistently implemented and often poorly interpreted. For example, a comprehensive review of education and psychology journals found that fewer than half of empirical studies reported any effect size, and only about 57% of those included an interpretation of its magnitude rather than just a numeric value (Sun, Pan, & Wang, 2010). Similar patterns were found across other psychology journals and even other disciplines such as health education and behaviour research (Barry et al., 2016; Fritz, Morris, & Richler, 2012, Farmus et al., 2023).

These patterns indicate that, despite longstanding methodological recommendations, effect sizes are not consistently used as tools for interpreting and communicating substantive significance in empirical research. This highlights the importance of developing instructional supports and tools that not only teach calculation of effect sizes but also emphasize interpretation and application in research contexts.

A Gap Between Reporting Standards and Statistical Education

Despite strong methodological recommendations from organizations such as the American Psychological Association to report and interpret effect sizes alongside significance tests (Wilkinson et al., 1999), empirical reviews indicate that both the reporting and, in particular, the interpretation of effect sizes remain inconsistent in psychological and educational research (Peng et al., 2013; Pek & Flora, 2018, Farmus et al., 2023).

As noted earlier, even when effect sizes are reported, they are not consistently interpreted in ways that convey practical significance. This gap in practice raises

questions about whether researchers themselves are fully equipped, conceptually or pedagogically, to interpret and communicate effect sizes in substantively meaningful ways, and whether statistical education sufficiently supports instructors in teaching these interpretive skills. This possibility aligns with documented omissions and misunderstandings of effect size concepts in pedagogical materials, indicating that limited coverage in foundational statistics instruction may contribute to both the superficial reporting of effect sizes and the challenges instructors face in moving beyond formula-based teaching toward interpretive competence (Morris, 2020).

However, instructional limitations may not fully account for problematic effect size reporting practices. Other factors, such as reviewer and editor preferences, may influence which effect size information authors ultimately include in published reports. Further, Sun et al. (2010) suggested that it was “reasonable for the authors to assume that the readers are aware of the definitions of the effect size measures, except for some uncommonly used measures” (p. 7). Such assumptions may contribute to variation in the extent to which effect sizes are defined, contextualized, and interpreted in published research, independent or researchers’ instructional backgrounds.

Moreover, the misalignment between reporting standards in the literature and classroom instruction may create a cyclical barrier. Instructors who themselves had limited exposure to effect size interpretation may feel constrained by time, curricular priorities, or perceived student readiness, and thus omit these topics or teach them superficially. Further, there is evidence that even teachers struggle to distinguish between inferential statistics and effect sizes correctly (Schmidt et al., 2023). Consequently, students graduate with incomplete skills in interpreting and

communicating effect sizes, which perpetuates the pattern of minimal or superficial reporting in the research they later conduct or evaluate. Recognizing and addressing this instructional gap was a central motivation for study 1 of the current research, which aimed to systematically identify the effect size techniques that instructors currently teach, omit, and the barriers underlying those decisions.

Student Difficulties and Misconceptions About Effect Sizes

Students frequently struggle with interpreting effect sizes, even when they can correctly calculate them, revealing a distinction between procedural competence and conceptual understanding. Common misconceptions include equating the magnitude of an effect with its practical or theoretical importance, misinterpreting benchmarks (e.g., small, medium, large) as universal thresholds, and failing to relate effect sizes to the research context or study design (Funder & Ozer, 2019). These difficulties are compounded by the emphasis in many introductory statistics courses on hypothesis testing and p-values, which can overshadow the importance of effect size interpretation. Empirical studies have documented that students often overestimate the meaning of statistically significant results and underestimate the importance of non-significant but substantively meaningful effects (Kühberger et al., 2015; Edelsbrunner & Thurn, 2024). This combination of procedural focus and conceptual misunderstanding underscores the need for instructional supports that explicitly target interpretation skills, providing students with opportunities to reason about effect sizes in context rather than solely computing formulas.

Instructional Challenges in Teaching Effect Size Interpretation

Teaching effect size interpretation presents multiple challenges for instructors, which may help explain both student difficulties and the inconsistent reporting observed in the literature. Research in statistics education demonstrates that instructors' attitudes toward concept-based or reform-oriented teaching significantly influence how deeply they engage students in conceptual reasoning rather than procedural calculation; perceived difficulty and confidence with these approaches are associated with less frequent adoption of interpretive instruction (Garfield & Ben-Zvi, 2008). In addition, instructors' confidence and familiarity with specific statistical concepts have been linked to their willingness and ability to teach those topics, suggesting that limited content knowledge may discourage instructors from covering interpretation-focused material (Harrell-Williams et al., 2015). Variability in faculty preparation further suggests that differences in academic background and experience can shape pedagogical practices in ways that affect emphasis on interpretive understanding, not just computation (Garfield & Ahlgren, 1988). Altogether, the statistics education literature identifies conceptual teaching challenges, instructor confidence, and attitudes as barriers that can constrain how effect size interpretation is addressed in introductory courses (Zieffler, Garfield, delMas, & Reading, 2008).

Technology-Supported Learning as a Potential Solution

In light of the instructional challenges associated with teaching effect size interpretation and the documented gaps in both reporting practices and classroom emphasis, researchers and educators have increasingly turned to interactive, technology supported learning environments as promising solutions for fostering deeper conceptual understanding (Freeman et al., 2014). Evidence from educational research

suggests that well designed digital tools and interactive resources can enhance cognitive engagement, support exploration of complex concepts, and provide immediate feedback that traditional instruction often lacks, contributing to improved learning outcomes across many domains (Hattie & Timperley, 2007). For example, interactive learning environments have been shown to yield positive effects on conceptual understanding and student engagement compared to conventional methods, particularly when they allow learners to manipulate variables and visualize abstract ideas dynamically (Rutten et al., 2012; Smetana & Bell, 2012). Moreover, studies in mathematics and statistics contexts indicate that integrating technological tools into instruction can increase conceptual understanding and active participation (Sokolowski & Willson, 2015; Chance et al., 2007; Garfield & Ben-Zvi, 2008), suggesting that interactive resources may be particularly well suited to address interpretation focused learning goals such as those involved in understanding effect sizes. These insights motivate the present investigation into the utility of newly developed interactive online resources designed to support effect size interpretation.

Overview of the Current Two-Study Research Program

This current research was designed as a two-study investigation addressing the persistent gap between statistical reporting standards in published research and the instructional practices used to teach effect size interpretation. Study 1 focused on assessing the instructional landscape: which effect size techniques instructors currently teach, which they omit, and the barriers influencing those decisions. Instructors were surveyed regarding their inclusion of eight well-established effect size techniques identified (see Cook et al. 2014, for details; Lakens & Caldwell, 2021). For techniques

not included in their courses, they were asked to provide rationales or identify barriers that shaped their teaching choices. A brief description of the eight techniques is provided in Table 1.

Table 1.

Summary of Effect Size Techniques

Technique	Description
Anchor based	Using an anchor or reference value which has known substantive meaning against which to compare the magnitude of differences. E.g., comparing patients who showed at least a certain level of improvement compared to patients who showed no improvement on a specific scale/instrument.
Distribution Based	Interpreting the size of an effect relative to the precision of the scale used to measure that effect.
Health-Economic (cost-benefit analysis)	Evaluating the costs involved in increasing an outcome by a certain magnitude. E.g., how much effort should a student exert using new study materials to increase their grade on a test by 5%.
Opinion-Seeking (expert opinion)	Consulting with experts within a substantive area to determine important or plausible magnitudes of an effect.

Pilot-Study	Conducting a smaller, though similar, study to produce an estimate of the effect size in question.
Review of Evidence Base (literature review or meta - analysis)	Conducting or using a meta-analysis to obtain or interpret an effect size estimate. Conducting a literature review for the same purpose as a meta-analysis.
Standardized effect sizes	Using published or recommended guidelines of standardized effect sizes to interpret the magnitude of an effect.
Simulation based effect sizes	Conducting a simulation to estimate ranges of plausible effect sizes. Simulation results could also be used to establish a reference point for interpreting the magnitude of an effect.

The findings from this study offer empirical insight into how instructional constraints, attitudes, and prior preparation contribute to gaps in students' conceptual understanding of effect sizes. These results also provide a foundation for a targeted intervention, highlighting areas where instructional support and resources could have the greatest impact.

Study 2 builds directly on the findings from Study 1 by evaluating the utility of newly developed interactive online resources designed to help students interpret effect sizes. These resources aim to address the conceptual and pedagogical challenges identified in study 1, including limited instructional time, and the difficulty of teaching abstract

concepts through conventional methods. By providing interactive, visual, and exploratory tools, study 2 investigates the quality and potential utility of technology-supported learning tools specifically designed to surmount these barriers which potentially lead to students' conceptual understanding and bridge the gap between statistical instruction and real-world research reporting practices.

The results from study 1 are essential for understanding how effect size instruction is currently implemented and why certain techniques are omitted, which in turn informs the design and focus of the study 2 intervention. Chapter 2 presents the methodology, survey development, participant characteristics, and findings from study 1, detailing the prevalence of effect size techniques in instruction and the reported barriers that contribute to their underutilization. This chapter establishes the empirical foundation for exploring instructional interventions, providing the context necessary to evaluate the potential impact of interactive online resources in study 2.

CHAPTER TWO:
ASSESSING THE PEDAGOGICAL LANDSCAPE

Purpose and Research Questions

Building on the conceptual and empirical background outlined in Chapter 1, study 1 of this research was designed to systematically examine the instructional practices and barriers associated with teaching effect size estimation and interpretation in psychology and related social science courses. Textbook analyses and survey evidence together highlight a systemic gap in effect size education: statistics textbooks often contain misunderstandings or omissions about effect sizes, and researchers report insufficient training, potentially leaving students without a strong conceptual foundation for understanding effect sizes (Morris, 2020; Collins & Watt, 2021). To better understand this gap, study 1 focused on identifying which of eight established effect size techniques instructors currently teach, which techniques they omit, and the reasons underlying these decisions. Specifically, the study addressed two central research questions: (1) Which effect size techniques are most commonly included or omitted in statistics instruction?, and (2) What barriers, constraints, or rationales do instructors report for not teaching certain techniques? Collecting these data was intended not only to map current instructional practices but also to provide an empirical foundation for study 2, where interactive online resources were developed and evaluated to target the specific instructional barriers or constraints identified in this survey.

Preregistration Plan

This study was conducted in accordance with open science best practices to enhance transparency and reproducibility. The research questions about predicted effect size technique prevalence rates (see Table 6 below), and analysis plan were

preregistered on the Open Science Framework (Adkins, 2024), with the time-stamped preregistration plan created on January 18, 2024. Although data collection for Study 1 had concluded prior to preregistration, no analyses were conducted until the preregistration plan was publicly posted, ensuring that analytic decisions were specified in advance of examining the data and reducing flexibility in analytic choices. The preregistered estimates for technique prevalence rates were based upon a combination of anecdotal observations accumulated over nearly a decade of experience as either a teaching assistant or course director, as well as the perceived complexity of each technique. The level of study at which each technique was taught was also considered when developing the preregistered prevalence rate estimates.

Survey responses were collected via the secure online platform formr.org, which incorporates features to protect participant anonymity. Data from Study 1 were briefly downloaded during collection solely to verify accurate recording and were deleted immediately; raw data were retained only after data collection concluded. To further safeguard confidentiality, raw response data are not publicly shared, given the small and specialized population. However, the preregistration materials, survey instrument, and analytic code are available upon request.

By integrating preregistration with robust data handling and reporting procedures, this study aligns with contemporary standards for transparency, ensuring that both methodological and ethical practices support replicable and credible research.

Methods

Participants

Ethical approval for the study was obtained from the Research Ethics Board at York University (Certificate # STU 2023-019) prior to data collection. The risk to participants was assessed to be minimal. A single inclusion criterion was applied: participants must have taught a statistics course within a social science department at a Canadian university within the previous five years. Participants could include tenured faculty, contract instructors, and graduate student instructors.

A purposive sampling strategy was used to identify potential participants. Canadian universities were identified using the Canadian Psychological Association's list of accredited programs. Faculty webpages were reviewed to identify instructors who taught statistics courses or who listed quantitative methods as a research interest. Institutional email addresses were compiled into a recruitment list (n = 86), and participants were contacted directly via email using blind carbon copy to protect recipient privacy. Identical recruitment emails were distributed through the CPA quantitative methods listserv and the York University quantitative methods listserv. The initial recruitment email was sent on April 3, 2023, followed by a reminder email on May 29, 2023. Participation was voluntary, and respondents self-selected into the study.

Study 1 Survey Instrument

The survey was organized into a series of sections designed to gather participant information about their educational background, current academic position, the level(s) of study at which they taught statistics, and questions focusing on their perspectives and understanding as it related to their teaching of effect size estimation and

interpretation. As part of the perspectives and understanding section, data on effect size misconceptions were collected using the questionnaire developed by Collins and Watt (2021). This questionnaire was designed to assess the extent to which respondents endorse common misconceptions about effect size interpretation. It was administered without modification and used with permission from the original authors. The items included in this questionnaire are presented in Table 5 of this current study.

The survey also employed conditional logic to ask follow-up questions where appropriate. Specifically, participants were presented with questions corresponding only to the educational levels they indicated teaching (e.g., undergraduate and/or graduate). Within each selected level, participants identified the effect size estimation and interpretation techniques they taught. For techniques that participants indicated they did not teach, additional questions explored perceived barriers to their inclusion within the participant's classroom. Further questions examined course-level characteristics associated with each educational level, including whether courses were required or elective, typical class sizes, and other instructional and curricular attributes relevant to the teaching of effect size concepts.

Standard demographic variables (e.g., age, sex assigned at birth, gender, ethnicity) were not collected. This decision was motivated by two considerations. First, there was no theoretical justification for expecting these variables to be associated with the study's research questions or objectives. Second, omission of detailed demographic questions was intended to further protect participant anonymity, as the combination of demographic characteristics could potentially render individuals identifiable within this relatively small and specialized population. Instead, limited background information was

collected to provide a general indication of participants' statistical training and instructional experience. This information included participants' highest degree obtained, the number of years since obtaining that degree, and their current academic appointment or instructional role (e.g., full professor, associate professor, assistant professor, lecturer, or graduate student instructor). The full survey can be found in Appendix A.

Data Collection Procedures

There was no formal stopping rule for data collection. Instead, data collection continued until no additional responses were forthcoming. Given the relatively small and specialized population of statistics instructors at Canadian universities, recruitment aimed to obtain responses from as many eligible instructors as possible rather than achieve a predetermined sample size. This approach is consistent with recommendations that sample size be justified in relation to the characteristics of the target population and the inferential goals of the study, including situations in which researchers seek to measure all or nearly all members of a finite population (Lakens, 2022). Because null hypothesis significance testing was not part of the preregistered analysis plan, the primary objective was to obtain a sample that was representative of Canadian university statistics instructors with respect to statistical training, teaching experience, and undergraduate and graduate teaching responsibilities, rather than to achieve a predetermined level of statistical power, consistent with the recommendations of Beribisky et al. (2025).

Following dissemination of the final reminder email, response rates were monitored. After a period of approximately two weeks elapsed with no additional responses, the survey was closed.

Results

A total of 125 individuals accessed the survey, of whom 32 provided consent and proceeded to the data collection phase. Most participants reported holding a PhD (n = 29), with one participant reporting a master's degree as their highest credential. Two participants did not complete any demographic or substantive survey items, and one additional participant completed only a subset of background information questions without responding to the effect size misconception items or other core sections of the survey. These three participants were excluded from all analyses, resulting in a final analytic sample of 29 instructors.

To contextualize instructors' reported instructional practices and perceived barriers, several descriptive indicators of training, professional experience, and teaching responsibilities were examined. These variables were not included as predictors in formal statistical models but are reported to characterize the sample and inform interpretation of subsequent findings.

Recency of Statistical Training

As a proximate indicator of the recency of formal statistical training, participants reported the number of years that had elapsed since obtaining their most recent academic degree. This measure was included to capture potential generational differences in statistical education, given substantial changes in statistical software,

computational accessibility, online training resources, and the emergence of the open science movement over the past two decades.

Response options were deliberately grouped into five-year intervals. This level of granularity was selected both to protect participant anonymity and to provide a meaningful overview of participants' training backgrounds and professional experience.

Participants were distributed across the full range of training recency. The largest proportion of respondents reported obtaining their most recent degree more than 20 years ago (n = 9). Eleven participants reported receiving their degree within the past 10 years, and the remaining nine participants reported completing their degree between 10 and 20 years ago. Table 2 presents the full frequency distribution.

Table 2.

Years Since Most Recent Degree

Years Since Last Degree	Count
Less than 5 years ago	4
More than 5 but less than 10 years ago	7
More than 10 but less than 15 years ago	5
More than 15 but less than 20 years ago	4
More than 20 years ago	9

Current Academic Position

A secondary indicator of professional experience was participants' current academic position. Although no formal statistical analyses were conducted using this variable, position titles provide a general rank ordering of experience within academic settings and help contextualize instructors' reported instructional practices and responsibilities.

Overall, the sample was composed primarily of established academics. Most respondents held tenure-track or tenured positions, with representation across assistant, associate, and full professor ranks. Smaller numbers of participants reported holding contract, lecturer, or other instructional or research positions. This distribution suggests that respondents generally brought substantial instructional and disciplinary experience to the study. Table 3 summarizes the distribution of academic positions.

Table 3.

Current Academic Position

Teaching Rank	Count
Full Professor	8
Associate Professor	7
Assistant Professor	7
Contract Faculty	2
Lecturer	3
Research or Teaching Employee without a PhD	1
Other	1

Level of Instruction

Participants also reported the instructional level(s) at which they taught statistics. Of the 29 respondents included in the analytic sample, 14 reported teaching statistics at both the undergraduate and graduate levels. Nine participants reported teaching only undergraduate statistics courses, and six reported teaching only graduate-level courses. Overall, 20 of the 29 respondents taught statistics at the graduate level.

This distribution suggests that a substantial proportion of instructors in the sample had experience teaching graduate-level statistical content. As such, limitations in instructor training may be less salient as a barrier to teaching effect size interpretation than originally anticipated in the preregistration plan. It is possible that instructors who teach across both undergraduate and graduate levels experience fewer instructional barriers due to greater exposure to statistical theory, applied methods, or curricular flexibility. These patterns are revisited in the interpretation of reported barriers.

Comfort Teaching Effect Size Interpretation and Estimation

Participants rated their comfort teaching effect size estimation and interpretation in university-level statistics courses using a 7-point Likert scale, with higher values indicating greater comfort. The average confidence rating was 6.38, indicating generally high self-reported comfort. The lowest rating reported by any participant was 5.

The majority of respondents reported being very or extremely comfortable teaching effect sizes, with 17 participants selecting the highest scale point. Table 4 presents the frequency distribution of comfort ratings.

Table 4.

Comfort Teaching Effect Size Interpretation and Estimation

Comfort Level	Count	Anchor Meaning
4	0	Neutral
5	6	—
6	6	—

Comfort Level	Count	Anchor Meaning
7	17	Extremely Comfortable

Overall, these results suggest that the sample consisted of instructors who were confident in their ability to teach effect size concepts, a finding that provides important context for interpreting both instructional omissions and reported barriers in subsequent analyses.

Perceived Importance of Effect Sizes

Consistent with the study's motivating assumptions, respondents expressed uniformly high endorsement of the importance of effect sizes in statistical practice and instruction. All participants rated effect sizes as either *Important* or *Very Important*. Specifically, five respondents selected *Important*, while the remaining 24 selected *Very Important*. No respondents indicated neutrality or disagreement. This finding establishes a strong normative consensus among instructors regarding the conceptual value of effect sizes and provides an important interpretive backdrop for subsequent analyses of instructional inclusion and omission.

Effect Size Understanding Among Instructors

Participants demonstrated high overall accuracy on the five-item effect size misconception assessment. Twenty-one respondents answered all five questions correctly, corresponding to 100% accuracy. Six respondents answered four of the five questions correctly (80% accuracy), one respondent answered three questions correctly (60% accuracy), and one respondent answered only one question correctly (20% accuracy).

At the item level, accuracy was consistently high, with no question falling below 79% correct. The lowest accuracy was observed for the statement assessing whether effect sizes express the magnitude of the influence one variable has on another (79.31%). This low accuracy could be related to the inclusion of the term “influence” which implies an established causal connection between variables which is not possible to determine from merely computing effect sizes. Several items (particularly those addressing common confusions between statistical significance, sample size, and effect size magnitude) exceeded 95% accuracy. These results suggest that, at the group level, instructors generally possessed a strong conceptual understanding of effect sizes, supporting the validity of their reported instructional practices and barrier endorsements. Item level accuracy can be found in Table 5.

Table 5.

Item level accuracy relating to effect size misconceptions

Statement	True Or False?	Correct Response Rate
Effect sizes express the magnitude of the influence one variable has on another	True	79.31%
A larger sample size will result in a larger (stronger) effect size	False	96.55 %
If you are doing high quality research, your aim is to find the largest effect sizes possible	False	89.66 %

Statement	True Or False?	Correct Response Rate
A statistically significant p-value corresponds to a medium or large effect size	False	96.55 %
A small effect size indicates that the null hypothesis should fail to be rejected	False	96.55 %

Patterns of Instructional Coverage in Undergraduate Courses

Instructional prevalence rates for effect size techniques in undergraduate statistics courses were compared to the predicted prevalence rates specified in the preregistration plan. Overall, observed inclusion rates were higher than anticipated for most techniques. The only technique for which the observed prevalence fell within the preregistered precision range was standardized effect sizes.

Table 6 presents the predicted prevalence rates estimates alongside observed prevalence rates for the 23 instructors from the initial sample of 29 who taught at least one undergraduate statistics course. Predicted prevalence estimates were generated for this subset of instructors. To examine whether observed prevalence rates differed as a function of teaching experience at a higher level, this subset was further divided into instructors who taught exclusively undergraduate courses (n = 9) and those who taught both undergraduate and graduate courses (n = 14). This further subdivision was conducted to assess whether experience teaching more advanced statistics was associated with differences in prevalence rates for undergraduate courses. The remaining 6 instructors from the initial sample taught exclusively graduate-level courses

and are reported separately using a similar analytical breakdown. Techniques such as anchor-based estimation, distribution-based methods, and pilot study approaches were taught at substantially higher rates than predicted. In contrast, techniques involving health economic approaches and opinion-seeking strategies remained relatively uncommon, though still more prevalent than preregistered estimates suggested.

Notably, instructors who taught both undergraduate and graduate courses tended to report higher inclusion rates for several techniques, particularly review-of-evidence approaches, than instructors who taught only undergraduate courses. These patterns suggest that instructional exposure across course levels may be associated with broader methodological coverage.

Table 6.

Undergraduate-Level Predicted Technique Prevalence Rates and Observed Prevalence Rates

Technique	**Predicted Estimate (%)**	**Instructors Who Taught at Least One UG Course n = 23 (%)**	**Instructors Who Taught Exclusively UG n = 9 (%)**	**Instructors Who Taught UG and Grad n = 14 (%)**
Anchor Based	0	52.17	55.56	50.00
Distribution Based	0	52.17	55.56	50.00
Health Economic	0	13.04	11.11	14.29
Opinion Seeking	50	30.43	44.44	21.43
Pilot Study	0	47.83	55.56	42.86
Review of Evidence Base	50	73.91	66.67	78.57
Simulation Based	50	30.43	33.33	28.57

Technique	Predicted Estimate (%)	Instructors Who Taught at Least One UG Course n = 23 (%)	Instructors Who Taught Exclusively UG n = 9 (%)	Instructors Who Taught UG and Grad n = 14 (%)
Standardised Effect Sizes	100	91.30	100.00	85.71

Barriers to Teaching Effect Size Techniques in Undergraduate Instruction

For each effect size technique that respondents indicated they did not teach, instructors were asked to identify perceived barriers to inclusion. Respondents were permitted to select multiple barriers and to provide open-ended comments.

Across techniques, the most frequently endorsed barrier was insufficient class time, followed by perceptions of technical complexity. Lack of familiarity with a technique was less commonly endorsed but emerged as a barrier for certain methods, particularly simulation-based approaches. These patterns were largely consistent across techniques, although some variability was observed. For example, simulation-based methods were most often described as overly complex, whereas health economic approaches were more frequently omitted due to time constraints.

The dominance of time-related constraints is notable given the high levels of instructor confidence and conceptual understanding observed elsewhere in the results. Together, these findings suggest that instructional omissions are driven more by curricular and structural limitations than by instructor knowledge deficits.

Patterns of Instructional Coverage in Graduate Courses

Instructional prevalence rates for effect size techniques in graduate-level statistics courses exhibited patterns broadly similar to those observed at the undergraduate level, though with several notable distinctions. One participant who reported teaching both

undergraduate and graduate statistics did not complete the graduate-level instructional coverage section; therefore, graduate-level analyses are based on 19 respondents.

As shown in Table 7, observed prevalence rates for most techniques exceeded the predicted estimates specified in the preregistration plan. Anchor-based and distribution-based approaches were reported by a majority of graduate instructors, despite low or zero predicted prevalence. Standardized effect sizes remained the most commonly taught technique, with nearly 90% of instructors reporting inclusion.

In contrast to undergraduate instruction, graduate-level teaching showed moderately higher inclusion rates for anchor-based, distribution-based, and review-of-evidence approaches. However, simulation-based techniques, predicted to be nearly ubiquitous at the graduate level, were taught by fewer than half of respondents. This discrepancy represents the largest deviation from preregistered expectations across both instructional levels.

Instructors who taught both undergraduate and graduate courses generally reported lower inclusion rates for several techniques compared to instructors who taught only graduate courses, particularly for anchor-based and simulation-based approaches. This pattern may reflect curricular trade-offs faced by instructors teaching across multiple levels or the need to maintain continuity across course sequences.

Table 7.

Graduate-Level Predicted Technique Prevalence Rates and Observed Prevalence Rates

Technique	Predicted Estimate (%)	Instructors Who Taught at Least One Grad Course n = 19 (%)	Instructors Who Taught Exclusively Grad n = 6 (%)	Instructors Who Taught UG and Grad n = 13 (%)
Anchor Based	25	63.16	83.33	53.85
Distribution Based	0	57.89	66.64	53.85
Health Economic	0	15.79	16.67	15.38
Opinion Seeking	50	42.11	50	38.46
Pilot Study	0	52.63	50	53.85
Review of Evidence Base	50	73.68	83.33	69.23
Simulation Based	100	36.84	16.67	46.15
Standardised Effect Sizes	100	89.47	83.33	92.31

Barriers to Teaching Effect Size Techniques in Graduate Instruction

Reported barriers to teaching effect size techniques at the graduate level closely resembled those observed for undergraduate instruction, with time constraints again emerging as the most frequently endorsed limitation. Despite the advanced nature of graduate coursework, instructors continued to cite insufficient class time as a primary reason for omitting specific techniques.

Perceived technical complexity was more salient at the graduate level for certain methods, particularly simulation-based approaches. Lack of familiarity was also more frequently endorsed for health economic techniques relative to other methods, suggesting that specialization rather than pedagogical difficulty may constrain inclusion of these approaches even in advanced courses. Notably, standardized effect sizes were

rarely omitted at the graduate level, and when they were, relatively few barriers were endorsed.

Open-ended “Other” responses were sparse and did not indicate systematic barriers beyond those captured by the structured response options.

Discussion

The primary aim of study 1 was to document current instructional practices related to effect size interpretation and estimation in social science statistics courses, identify techniques that were omitted within classrooms, and understand the barriers that shape these instructional decisions. Taken together, the results provide a nuanced picture of a field in which effect sizes are widely valued, instructors are generally confident and conceptually competent, yet curricular constraints continue to limit the breadth of techniques presented to students.

Instructor Values and Conceptual Readiness

Across the sample, there was clear consensus regarding the importance of teaching effect size techniques in statistics instruction. This shared valuation was reinforced by strong performance on the effect size misconception assessment and high self-reported comfort teaching effect size estimation and interpretation. These converging indicators suggest that, at least within this sample, instructors possess the conceptual foundations necessary to teach effect sizes effectively.

Instructional Coverage Relative to Preregistered Expectations

Observed instructional prevalence rates diverged from preregistered predictions, particularly at the undergraduate level. With the exception of standardized effect sizes, most techniques were taught at higher rates than anticipated. This finding suggests that

effect size instruction may be more widespread than informal curricular accounts imply, especially for techniques that align closely with traditional statistical workflows.

However, this pattern was not uniform across techniques. Opinion-seeking and simulation-based approaches remained relatively uncommon in undergraduate instruction, and simulation-based methods, in particular, attracted the largest number and diversity of reported barriers. These findings point to a distinction between conceptual endorsement of effect sizes and the practical feasibility of teaching more resource-intensive or technically demanding approaches.

At the graduate level, instructional coverage generally fell below preregistered expectations, with the notable exceptions of opinion-seeking, simulation-based, and standardized effect size techniques. The lower-than-expected prevalence of simulation-based methods at the graduate level was especially striking, given their presumed alignment with advanced statistical training. The substantial discrepancy between instructors who taught only graduate courses and those who taught across both undergraduate and graduate levels suggests that teaching context and workload distribution may meaningfully shape curricular decisions, even in advanced courses.

Barriers as Structural Rather Than Conceptual

Across both instructional levels, time constraints emerged as the most frequently cited barrier to teaching effect size techniques. Perceived technical complexity (particularly for simulation-based approaches) was also commonly endorsed, while lack of familiarity played a more limited role. Health economic approaches were consistently the most under-utilized technique. Instructors across all levels of instruction most commonly cited insufficient class time as a barrier to inclusion, followed by a lack of

familiarity with the technique to teach at a particular level. These findings suggest that health economic approaches may be excluded primarily because of curricular constraints and limited instructor expertise rather than disagreement with their value.

Taken together, these findings indicate that instructional omissions are driven primarily by structural and logistical factors rather than by instructor misunderstanding or resistance. This distinction is critical for intervention design: efforts to improve effect size instruction are more likely to succeed if they focus on reducing instructional burden, increasing accessibility, and integrating techniques seamlessly into existing curricula.

Scope of Instructional Impact

Although the sample size was modest, the potential instructional impact of the participating instructors is substantial. All respondents had taught statistics within the past five years, many taught multiple required courses, and several reported large class sizes. To estimate the number of students potentially exposed to these instructors' teaching practices, the respondents were asked to report their approximate minimum and maximum class enrollments, and the number of required and non-required introductory statistics classes they taught during the past 5 years. A conservative instructional impact estimate was calculated by taking the minimum reported class size for each instructor and multiplying it by each instructors reported number of required introductory statistics classes. Each of these values was summed to arrive at the final impact estimate for the sample. This resulted in an impact estimate of 2,482 students, which is about 496 students per year for the previous 5 years. This approach intentionally produced a cautious estimate that likely understates the total number of students reached. An upper bound impact was also estimated by multiplying each

maximum class size by the total number of introductory statistics classes for each instructor. This resulted in an impact estimate of 7,502 students which is about 1500 students per year for the previous 5 years. For example, if an instructor reported teaching 5 courses (2 of which were required) with class sizes ranging from 25-150 students, then this instructor's individual conservative impact estimate would be 25*2=50 students. This instructor's upper bound would be 5*150=750 students, though this estimate might be somewhat inflated if the same students took all their courses from the same instructor. Even when considering midpoint estimates between these bounds, the findings suggest a considerable instructional reach for a relatively small number of instructors.

These estimates underscore the practical significance of instructional decisions about effect size techniques. Small changes in instructional coverage—particularly in required courses—have the potential to affect thousands of students' statistical reasoning and interpretive skills. As such, improving the accessibility and feasibility of teaching under-represented effect size techniques may yield outsized benefits relative to the scale of intervention required.

Implications for Study 2 and Instructional Design

The findings from study 1 suggest that the primary challenge in effect size education is not persuading instructors of its importance or remedying widespread misconceptions, but rather supporting instructors in navigating time constraints and pedagogical complexity. These results directly motivate the design goals of study 2, which aim to develop instructional resources that are time-efficient, conceptually clear, and adaptable across instructional levels.

By grounding intervention design in empirically identified barriers and existing instructional practices, this study advances a pragmatic approach to statistical reform, one that leverages instructor readiness while addressing the structural realities of contemporary statistics education.

CHAPTER THREE:
HISTORICAL AND THEORETICAL FRAMING

Introduction

Study 1 demonstrated that instructors generally value effect size estimation and interpretation, possess the conceptual knowledge required to teach these techniques, and recognise their importance within statistics education. However, instructional coverage varied considerably across techniques, with time constraints, curricular demands, and perceived complexity emerging as the primary barriers to broader implementation. These findings suggest that improving effect size education may depend less on changing instructor attitudes and more on providing accessible instructional resources that reduce implementation burden while supporting conceptual learning.

An interactive website was designed to support conceptual understanding of effect size estimation and interpretation techniques. This website was intended to function as a flexible, self-directed learning environment that could supplement instruction and increase exposure to underrepresented effect size techniques.

This chapter outlines the theoretical, pedagogical, and technological foundations guiding the development of the interactive website. It begins by situating the work within broader developments in interactive statistics education, before outlining the statistical learning principles and instructional design frameworks that informed the learning activities. It then describes human-centred and behavioural design considerations and concludes by showing how findings from Study 1 were translated into concrete design requirements and implemented within the website.

Evolution of Interactive Online Resources in Statistics Education

The design of the website reflects broader developments in statistics education and educational technology. Over time, instructional priorities in statistics education have shifted from computational proficiency toward statistical reasoning and interpretation, reflecting broader reform movements that emphasise conceptual understanding, statistical literacy, and thinking with data. This shift has been accompanied by increasing recognition that learners benefit from active engagement with data and representations rather than passive reception of procedural instruction (Carver et al., 2016; Garfield & Ben-Zvi, 2009; Franklin et al., 2007).

Interactive technologies have played a central role in enabling this transition, providing opportunities for learners to manipulate data, explore variability, and visualize statistical relationships dynamically. This section outlines this progression to situate the present website development within established pedagogical and technological developments in statistics education.

From Computational Instruction to Statistical Reasoning

Historically, statistics instruction emphasised procedural competence, including hand calculations, formula memorisation, and the application of fixed analytical procedures (Lovett & Greenhouse, 2000; Cobb, 2007; Tishkovskaya & Lancaster, 2012). This approach has been characterised in the statistics education literature as focusing on computational proficiency at the expense of conceptual understanding and statistical reasoning (Garfield & Ben-Zvi, 2009). While these approaches supported computational accuracy, they often provided limited opportunities for conceptual understanding or interpretation.

Gigerenzer (2004) critiques this tradition as producing “mindless statistics,” where learners are able to generate statistical outputs without understanding their meaning or implications. In response, statistics education has increasingly prioritised statistical literacy, thinking, and reasoning (Garfield & Ben-Zvi, 2007).

This shift is also reflected in the GAISE framework, which emphasises conceptual understanding, active learning, and engagement with authentic data as central instructional goals (Carver et al., 2016). Within this perspective, statistical education is increasingly understood as a reasoning-based discipline rather than a computational one, with emphasis placed on interpretation and decision-making in context (Ben-Zvi & Garfield, 2008). This reconceptualization of statistical learning provides the pedagogical foundation for interactive approaches that support exploration and interpretation rather than procedural repetition.

Emergence of Technology-Supported Statistics Learning

As statistical education shifted toward reasoning-based goals, educational technology became increasingly important in supporting these aims. Early educational technologies in statistics were largely designed as tools for computation and data analysis, reducing the burden of manual calculations, whereas later educational software was developed specifically to support students' conceptual understanding and statistical reasoning (McNamara, 2019; Zavez & Harel, 2025).

Subsequent developments introduced dynamic and exploratory environments that allowed learners to interact with statistical representations. Garfield and Ben-Zvi’s (2005) work on reasoning about variability highlighted the importance of repeated

exposure to statistical phenomena and dynamic visualisation for developing conceptual understanding.

Chance et al. (2007) further argued that technology can support learning through simulation, visualization, and exploratory interaction with data. These tools shift learners from passive recipients of information to active investigators of statistical relationships. This progression has evolved from static textbooks to statistical software, then to applets and web-based simulations, and most recently to interactive platforms such as R Shiny and JavaScript-based frameworks. These contemporary tools enable building fully integrated environments in which learners can explore, manipulate, and receive immediate feedback on statistical concepts (Lu, 2023; Zavez & Harel, 2025).

Importantly, these developments also address practical constraints in education by enabling asynchronous access and repeated engagement with material, making them particularly suitable for contexts where instructional time is limited.

Interactive Learning as a Response to Persistent Statistical Misconceptions

Despite advances in statistics instruction and the use of interactive technologies, learners continue to struggle with core statistical concepts, particularly variability, uncertainty, and interpretation (Garfield & Ben-Zvi, 2009; Zieffler, Garfield, Alt, Dupuis, Holleque, & Chang, 2008). These difficulties are well documented in the literature and often persist even when students demonstrate procedural competence in statistical calculations (delMas et al., 2007).

The Statistical Reasoning Learning Environment framework emphasises that statistical understanding develops through active engagement, exploration, and reflection rather than passive instruction (Garfield & Ben-Zvi, 2009). Within this model,

learners construct understanding by testing ideas, observing outcomes, and refining interpretations through experience.

Interactive technologies are well aligned with this approach because they allow learners to manipulate variables, explore alternative scenarios, and receive immediate feedback. These features support iterative learning processes that are difficult to achieve in static instructional formats. As a result, interactive environments are poised to be effective tools for addressing persistent misconceptions in statistics education by enabling experiential and exploratory learning.

Contemporary Principles for Interactive Educational Resources

As educational technologies have developed, research has increasingly emphasised that learning outcomes depend not only on instructional content but also on the usability and user experience of the technologies through which learning is delivered (Vlachogianni & Tselios, 2022).

Key characteristics associated with effective educational websites include clarity, simplicity, readability, intuitive navigation, and purposeful interaction (Garett et al., 2016; Schneider et al., 2012). These features are particularly important in statistics education, where cognitive load is often high and learners may already struggle with abstraction.

A related distinction is made by McNamara (2019), who differentiates tools designed for statistical analysis from those designed for statistical learning. While analytical tools focus on computation, learning tools must prioritise conceptual development and interpretation. Consequently, learning tools should adhere to a different design process given that such tools are designed for decidedly different goals relative to statistical analysis tools.

Human-centred design principles further emphasise the importance of designing technologies around user needs and context of use, with iterative evaluation throughout development (International Organization for Standardization 9241-210, 2019). Such design principles are paramount to successfully developing statistical learning tools that are not only pedagogically sound but also usable in practice. Together, these perspectives highlight the importance of integrating usability and user experience in the design of educational technologies.

Implications for Effect Size Education

The developments outlined above suggest that interactive learning environments are particularly well suited to supporting effect size education. Effect sizes require interpretation of magnitude, variability, and context making them inherently conceptual rather than procedural.

Although effect size reporting is widely recognised as important in psychological research and evidence synthesis (Farmus et al., 2023), instructional coverage remains inconsistent. Study 1 of the present research supported this finding; specifically, although instructors value and understand effect sizes, their implementation is constrained by time, curricular structure, and the perceived complexity of certain techniques. Furthermore, a recent literature review on the use of computing tools for teaching statistical concepts broadly categorised articles according to the statistical concepts addressed, and very few articles (~4.9%) focused on effect sizes or related techniques (Zavez & Harel, 2025).

Statistics Education Framework

This section outlines the statistical education principles that directly informed the design of the learning activities. While the previous section established the broader rationale for interactive learning, this section focuses on the specific conceptual foundations underlying effect size instruction within the intervention.

Statistical Reasoning as the Primary Learning Goal

This website's development prioritised statistical reasoning over procedural computation as the primary learning goal. Although effect sizes are calculated using formulas, their educational value lies in interpretation and contextual meaning. In service of this learning goal, a design decision was made to hide complex computations and calculations from users and perform them behind-the-scenes. This design decision was made to not only mitigate the potentially high cognitive load for users, but to also help guide learners' attention by focusing them on specific aspects which are pertinent for specific lesson learning objectives.

Accordingly, the website's interactive activities focused on explicating the underlying rationale for specific effect size techniques to teach learners how to transition from numerical outputs to substantive conclusions while also considering context.

Data-Based Learning Through Interaction

The website also adopted a data-based inquiry approach in which learners engage directly with datasets rather than abstract definitions. This approach aligns with calls for data-centred instruction in statistics education (Carver & Stephens, 2014), which emphasise exploration of distributions, patterns, and relationships.

Within the website, learners manipulated data, compared groups, and observed how structural changes affect effect size estimates. These interactions supported active

exploration and grounded interpretation in empirical evidence rather than abstract rules. As an additional enhancement to promote repeated usage and subsequent learning, dynamic and randomly generated data sets were used rather than static sets of data. Learners could retry exercises with new data sets with the click of a button and each time a webpage is visited, the data are automatically regenerated.

Instructional Design Framework

This section translates statistical education principles into instructional structure, describing how learning was organised within the website to support conceptual development.

Constructivist Learning Principles

The website development was grounded in constructivist learning theory, which posits that learners actively construct knowledge through interaction with experiences, prior knowledge, and social contexts, rather than passively receiving information (Piaget, 1953; Loyens & Gijbels, 2008; Garfield & Ben-Zvi, 2009). In statistics education, this is particularly important because understanding depends on interpretation and sense-making rather than memorisation.

Activities were therefore designed to encourage exploration and reflection. Learners were prompted to test expectations, observe outcomes, and revise understanding through interaction with statistical representations. This aligns with evidence from educational psychology showing that active cognitive engagement improves conceptual learning and retention (Chernikova et al., 2020).

Active Learning and Immediate Feedback

Consistent with GAISE (Carver et al., 2016), the intervention incorporated active learning supported by immediate feedback. As learners engaged in tasks requiring effect size estimation and interpretation, immediate feedback designed to reinforce correct understanding, and knowledge synthesis was provided.

Research indicates that active participation combined with feedback improves learning outcomes across educational contexts (Lysakowski & Walberg, 1982), and that scaffolded feedback in digital environments supports conceptual development in complex domains (Chernikova et al., 2020).

Scaffolding Statistical Interpretation

Scaffolding reduces cognitive load and supports the incremental development of complex reasoning skills (Doo, Bonk, & Heo, 2020). In the development of the website, interactive activities were designed to elicit and build upon learners' intuitive judgments, providing a foundation upon which more formal statistical concepts could be developed. Through a scaffolded sequence of tasks, learners progressively refined these intuitions by incorporating additional contextual information and empirical evidence. Simulation and visualisation tools further supported this process by making abstract statistical relationships observable, enabling learners to compare their expectations with empirical outcomes and ultimately synthesise multiple sources of information to arrive at deeper conceptual understanding (Tintle et al., 2015).

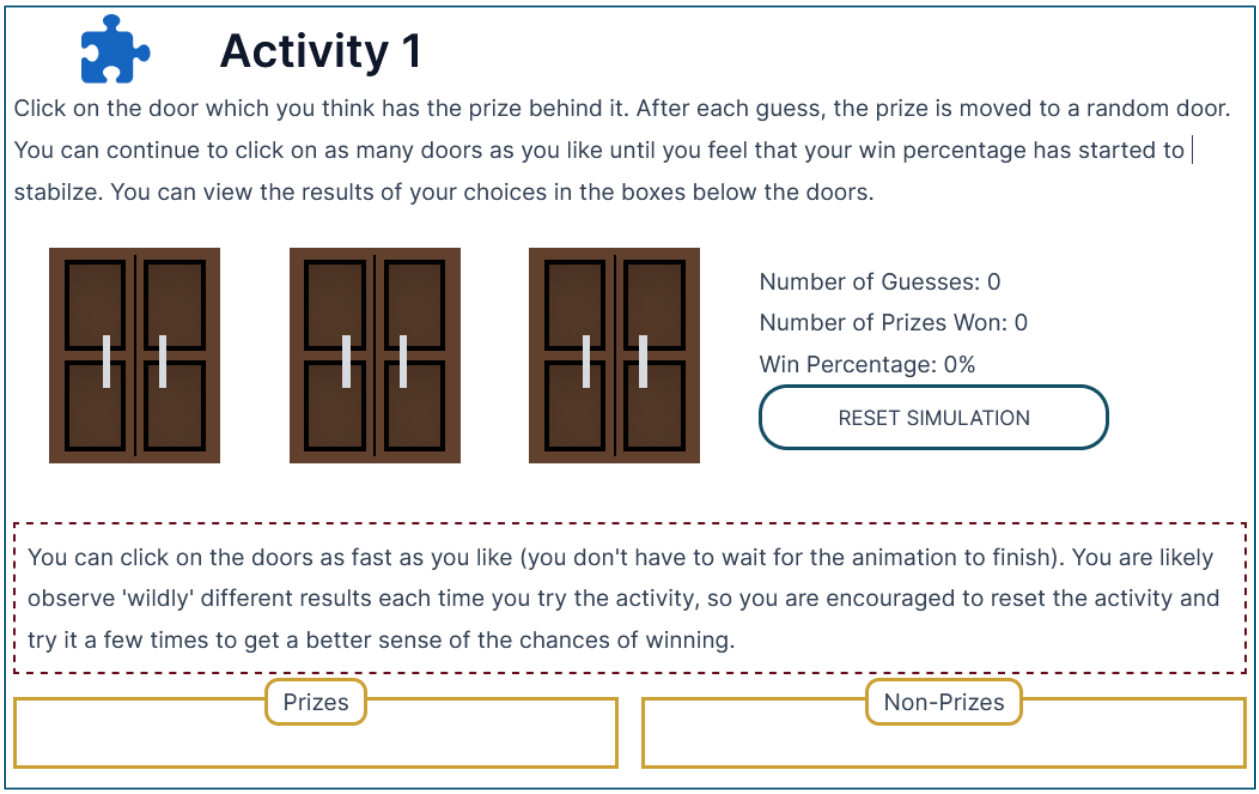
A representative example of this design approach is the lesson on simulation-based effect size estimation and interpretation. The lesson begins with an introduction to the Monty Hall problem, a well-known probability puzzle in which a prize is hidden behind one of three doors. Learners initially select a door, after which the game host

reveals one of the remaining non-prize doors and offers the opportunity to either retain the original selection or switch to the remaining unopened door. Before introducing the full problem, however, learners are presented with a simplified version of the game.

In this initial activity, the game host does not reveal any additional doors. Instead, learners simply select a door and observe whether a prize is obtained. The activity (see Figure 1) was intentionally designed to establish intuitive expectations regarding random selection, uniform distributions, and basic event probabilities without relying on formal statistical terminology.

Figure 1.

Screenshot of Simulation-Based Interactive Activity

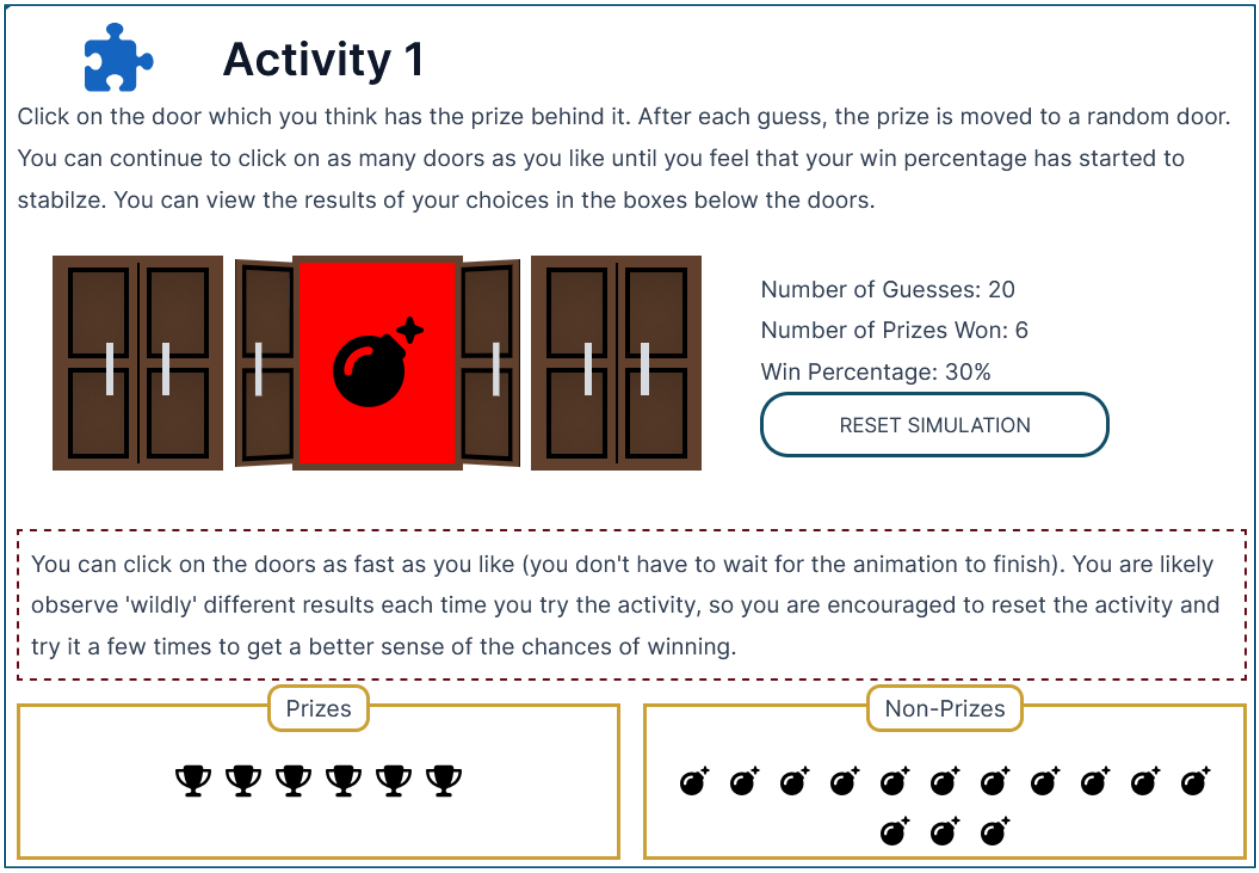


As learners continue making selections, the webpage updates dynamically to display cumulative results in real time (see Figure 2). Outcomes are presented through

both numerical summaries and graphical representations using collections of icons. This immediate feedback allows learners to observe emerging patterns across repeated trials and compare these patterns with their initial expectations. The activity therefore serves as an intuitive introduction to long-run relative frequency and probabilistic reasoning.

Figure 2.

Updated Screenshot of Simulation-Based Interactive Activity

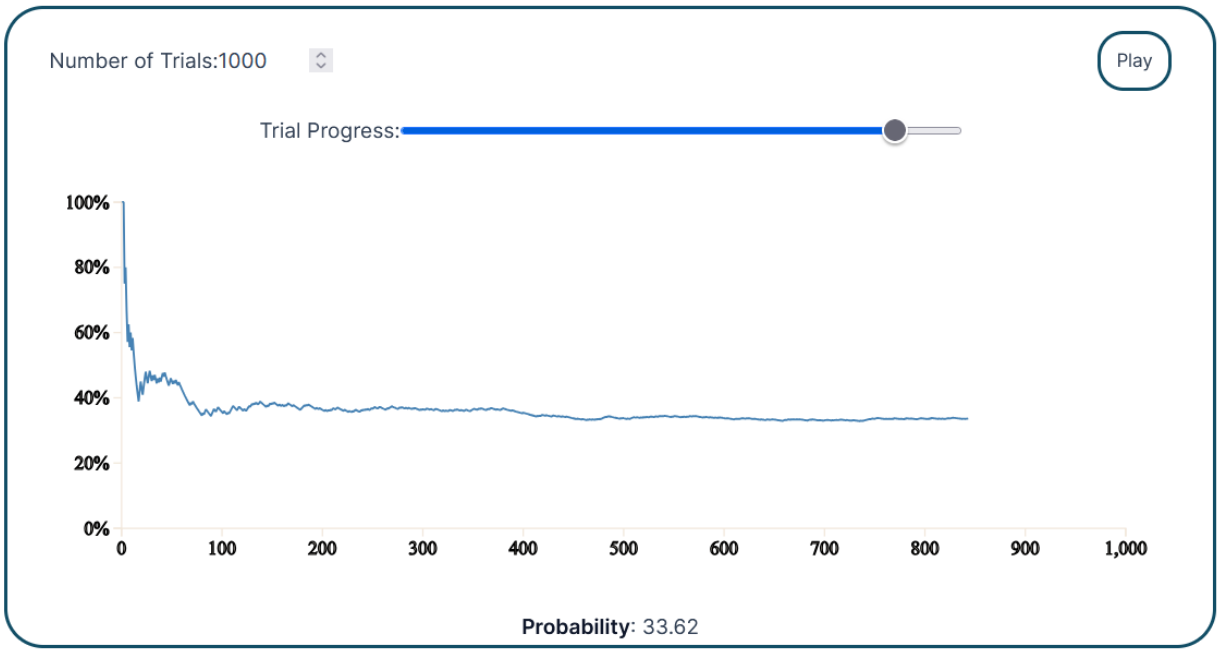


The subsequent lesson page introduces a graphical simulation designed to evaluate learners' intuitions regarding the long-run probability of success in the simplified game (see Figure 3). The simulation explicitly demonstrates that, although random fluctuations occur across smaller samples, the probability of obtaining a prize

converges toward approximately one-third over repeated trials. The graphical display updates continuously as the simulation runs within the browser, providing a visual representation of sampling variability and convergence. Participants can also manipulate the number of simulation trials, ranging from 10 to 10,000, thereby observing the impact of sample size on the stability of probabilistic estimates.

Figure 3.

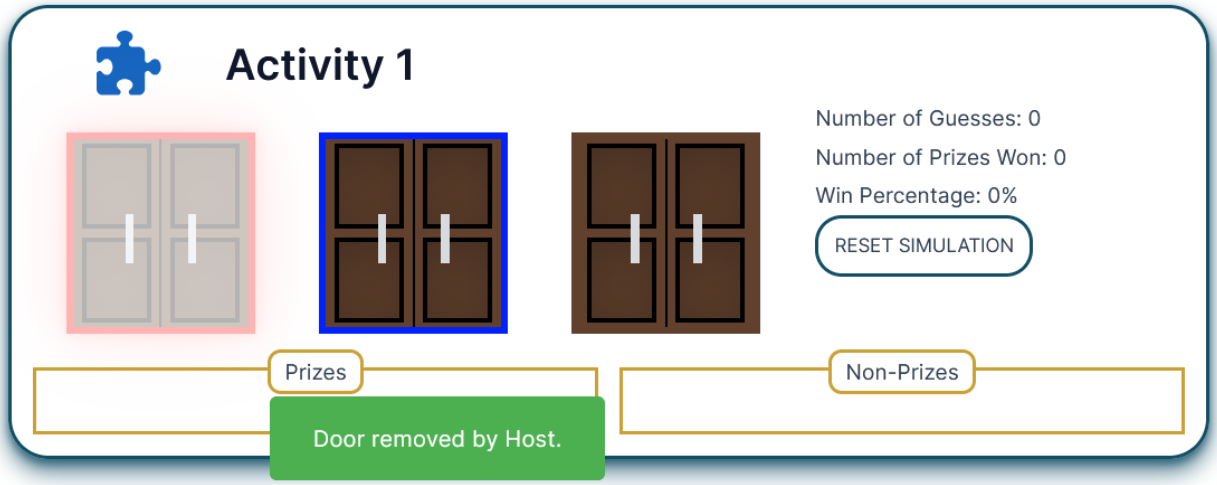
Screenshot of Simulation-Based Interactive Activity for Long-Run Probabilities



Following this introductory sequence, the lesson presents the full Monty Hall problem. In this version, the host removes a non-prize door after the participant's initial selection and then offers the choice of retaining the original door or switching to the remaining unopened door (see Figure 4). Participants are invited to explore different decision strategies and evaluate whether one strategy yields a greater probability of success over repeated trials.

Figure 4.

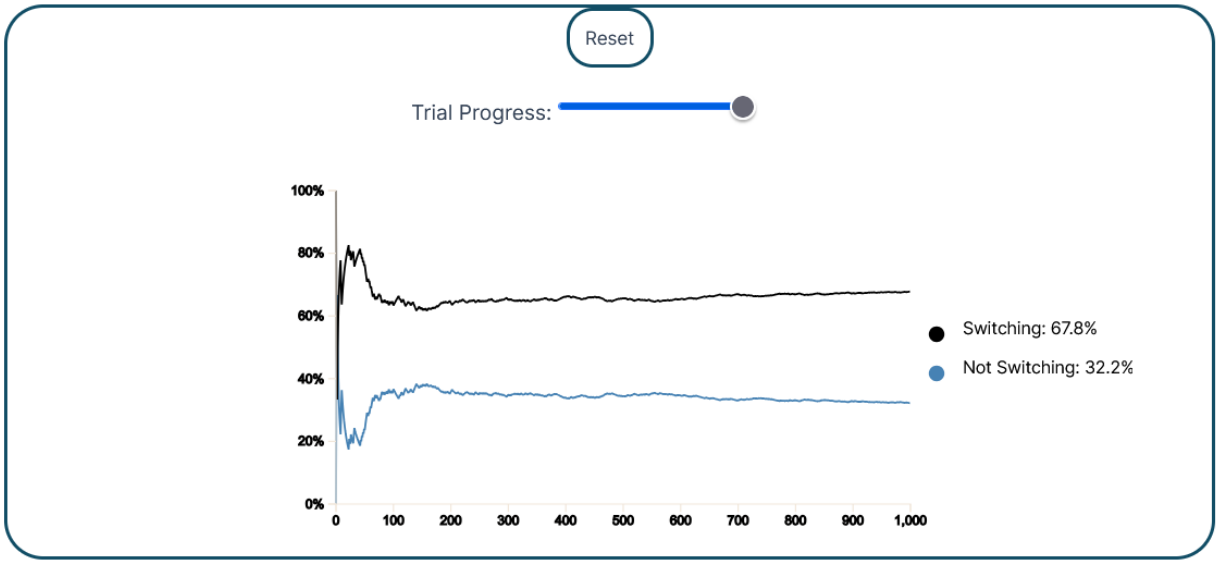
Screenshot of The Full Monty Hall Simulation



The next activity extends this exploration through an animated comparison of two strategies: consistently retaining the original door choice or consistently switching after a door has been removed. Building on learners' previously developed intuitions regarding probability, the activity illustrates how seemingly minor changes in decision-making behaviour can produce substantial differences in long-run outcomes (see Figure 5). Specifically, the visualisation highlights the increase in winning probability from approximately 33% when retaining the original choice to approximately 66% when switching. This comparison was intended to introduce learners to the idea that the practical significance of an outcome is inherently comparative, foreshadowing later discussions of effect size interpretation. In other words, the magnitude of an effect becomes meaningful only relative to an appropriate reference point or comparison condition.

Figure 5.

Screenshot of the Long-Run Probabilities for Different Door Selection Strategies



The lesson sequence concludes with an activity titled “Build-A-Simulation”, which centres on the guiding question, "What if?" At this stage, learners are provided with direct control over key simulation parameters, including the number of doors in the game and the number of doors removed by the host (see Figure 6 and Figure 7). As parameters are modified, the simulation updates in real time, visually representing changes to the structure of the problem and the resulting probabilities. This final activity was designed to extend and consolidate previous learning by encouraging learners to test hypotheses, explore alternative scenarios, and generalise principles beyond the original problem context. By shifting learners from observing simulations to constructing and modifying them, the activity supports deeper conceptual understanding and promotes the transfer of probabilistic reasoning to novel situations.

Figure 6.

Screenshot of the Build-A-Simulation Activity

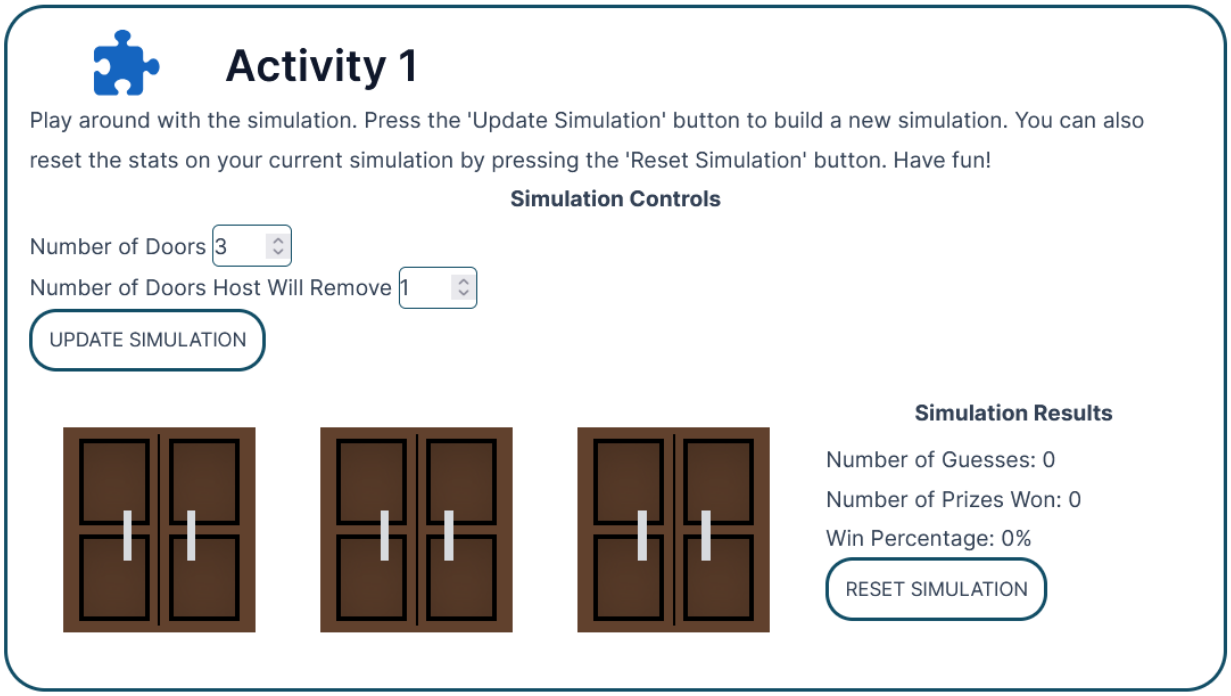
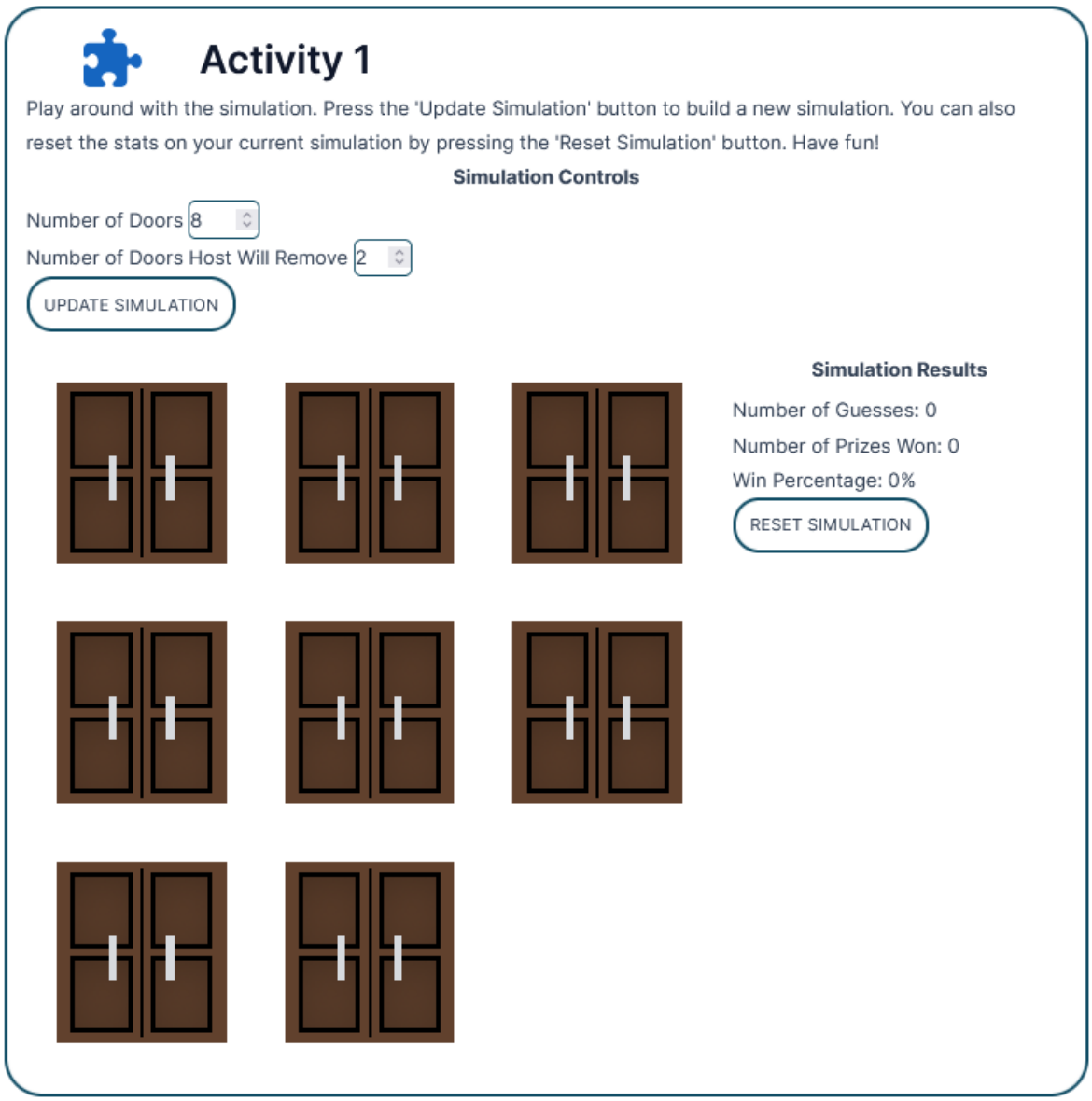


Figure 7.

Screenshot of a Customized Monte Hall Simulation Activity



Translating Study 1 Findings Into Design Requirements

The findings from Study 1, taken in conjunction with existing literature discussed above, informed the development of the website through a systematic translation of identified instructional challenges into specific design requirements. Rather than serving solely as descriptive findings, participants' reported barriers and needs were used to

guide key design decisions, ensuring that the resulting educational resource directly addressed issues encountered by instructors in practice. Table 8 summarises the relationship between the principal findings from Study 1, the corresponding design requirements, and the features implemented within the website.

Table 8.

Translation of Study 1 Findings and Literature Themes to Website Design Features

Finding	**Requirement**	**Feature**
Limited time	Self-paced structure	Modular learning units
Need for conceptual understanding	Interpretation focus	Interactive visualisations, scaffolded lesson plans
Technique complexity	Accessibility	Jargon-free activities, hidden computations and formulae
Misconceptions	Scaffolding	Guided feedback tasks
Uneven coverage	Independence	Standalone modules
Equal access	No-cost solution	Non-proprietary software, and open-source solutions

Several themes were particularly influential in shaping the design and development of the website. First, in Study 1, instructors frequently reported constraints related to instructional time, suggesting that extensive additions to existing curricula would be difficult to implement. To address this challenge, the website was structured as a series of self-paced, modular learning units that could be completed independently or incorporated selectively into existing courses.

Second, existing literature emphasised the importance of developing conceptual understanding rather than simply teaching computational procedures (Garfield & Ben-Zvi, 2007; Ben-Zvi & Garfield, 2008; Garfield & Ben-Zvi, 2009). Consequently, the website prioritised interpretation and reasoning through the use of interactive visualisations, simulations, and scaffolded lesson sequences. These features were intended to support learners in understanding the meaning and practical significance of effect size estimates rather than focusing exclusively on their calculation.

Third, the limited availability of dedicated teaching resources for effect sizes highlighted the need for a readily accessible instructional tool. This requirement motivated the development of a web-based platform that could be accessed without specialised software or institutional resources, thereby reducing barriers to adoption for both instructors and learners.

Finally, the uneven coverage of effect size techniques across courses and programs suggested that learners may enter instructional contexts with varying levels of prior knowledge. As a result, the intervention was structured as a collection of standalone modules that could be completed independently, allowing learners to engage with specific topics according to their needs without requiring completion of the entire curriculum.

Overall, this design process demonstrates a clear line of evidence from the empirical findings of Study 1 and the existing literature to the features incorporated within the website design and development. By grounding design decisions in instructors' reported experiences and instructional needs, the resulting resource was

intended to address identified barriers while supporting more effective teaching and learning of effect size concepts.

CHAPTER FOUR:
ASSESSING ONLINE INTERACTIVITIES

Introduction

Background

Building directly on the findings of study 1, study 2 moves from documenting instructional practices to evaluating a potential structural solution: purpose-built online activities designed to support the teaching of effect size estimation and interpretation. Study 1 demonstrated that instructors value effect size instruction and possess the conceptual readiness to teach it effectively, yet face persistent barriers related to time constraints, perceived technical complexity, and curricular crowding. These findings suggest that meaningful reform is less likely to emerge from persuasion alone and more likely to depend on resources that reduce instructional burden while preserving conceptual rigor. In response, study 2 gathered systematic feedback from statistics instructors regarding a set of online activities developed to teach effect size estimation and interpretation in ways that are adaptable across undergraduate and graduate contexts.

Crucially, the design of these online activities was informed not only by the structural barriers identified in study 1, but also by established principles of effective web-based instructional design (e.g., Garett, Chiu, Zhang, & Young, 2016). The website was evaluated across seven core design features: navigation, graphical representation, organization, content utility, purpose clarity, simplicity, and readability. Navigation emphasized consistent and user-controlled movement across sections, allowing instructors and students to access content efficiently without cognitive friction. Graphical representation focused on visual clarity and minimal overload, including appropriate

typography, image resolution, effective white space, and optimized loading times—features particularly important for simulation-based demonstrations. Organizational structure ensured that headings, labels, and content sequencing were logical, consistent, and pedagogically transparent. Content utility centered on the provision of accurate, current, and relevant material that motivates further exploration, aligning directly with instructors' desire for high-quality, classroom-ready resources.

In addition, purpose clarity was prioritized so that each activity explicitly communicated its instructional objective and unique contribution to effect size understanding. Simplicity guided interface decisions, emphasizing ease of use, intuitive features, device optimization, and uncluttered layout; each of these design elements intended to reduce adoption barriers for time-constrained instructors and students alike.

Finally, readability considerations ensured that written explanations were clear, well-edited, and accessible to a broad range of learners, from novices encountering effect sizes for the first time to more advanced students refining interpretation skills. By systematically integrating these seven design dimensions, study 2 not only examines instructor perceptions of the activities' pedagogical value, but also formally assess the delivery method, activity design, and overall user experience.

Technologies

Beyond pedagogical design considerations, the technological infrastructure underlying the online activities was intentionally selected to align with the structural concerns identified in study 1. Because instructors cited time constraints and technical complexity as primary barriers to expanding effect size instruction, the development process prioritized tools that are stable, widely accessible, low-cost, and scalable

across institutional contexts. The goal was not merely to produce interactive demonstrations, but to do so in a manner that minimizes implementation burden and maximizes adaptability ultimately leading to adoptability.

The interactive visualizations were developed using D3 (D3v7, Data-Driven Documents; Bostock, 2021), a JavaScript library that enables dynamic, data-responsive graphical representations within standard web browsers. D3 allows statistical quantities to update in real time as users manipulate parameters. With proper care and design, these interactive web components can be seamlessly integrated within webpages, leaving the technical parts behind the “curtain” and lead to more intuitive use cases. Importantly, D3-based elements operate natively within modern browsers, eliminating the need for specialized software installation and thereby lowering technical barriers for both instructors and students. A caveat to this JavaScript library is the steep learning curve for developing new resources.

All statistical computations and data generation procedures were developed in R using the RStudio integrated development environment (Posit team, 2024). R was selected due to its widespread adoption in social science research and graduate-level statistics training, as well as its open-source status and extensive community support. The use of R facilitated transparent implementation of effect size formulas, simulation routines, and reproducible workflows. Moreover, structuring the computational backbone in R supports future expansion, as instructors and researchers familiar with the language can inspect, modify, or extend the underlying code base without reliance on proprietary systems.

Critically, once compiled and deployed, the activities function entirely on the client side within the user's browser. After the initial page load, no ongoing server–client communication is required for simulations, parameter manipulation, or visual updates. This architecture offers several advantages directly relevant to the barriers identified in study 1. First, it reduces institutional IT constraints, as the activities do not require server hosting infrastructure or database management. Second, it enhances reliability in classroom settings by minimizing latency and dependency on external computational resources. Third, it improves the website's scalability by minimizing the impact of concurrent users on server workload. As a result, the system can accommodate higher levels of usage without compromising performance, reducing the likelihood of latency or server failures that could negatively affect the user experience.

All technologies employed in the development process (D3, R, and standard web technologies such as HTML, CSS, and JavaScript) are free, open-source, and have cross-browser-compatibility. This open technological ecosystem invites collaboration, iterative improvement, and adaptation across institutions. Instructors are not positioned as passive consumers of a fixed instructional product; rather, the design permits modification, redistribution, and expansion. In this way, the technological choices underlying the online activities mirror the broader aims of study 2: to create accessible, sustainable, and structurally feasible tools that support comprehensive effect size instruction without increasing instructor burden.

Methods

Participants

Ethical approval for this study was obtained from the Research Ethics Board at York University (Certificate # STU 2023-019) via an amendment to the original ethics application. The risk to participants was assessed to be minimal. A single inclusion criterion was applied: participants must have taught a statistics course within a social science department at a Canadian university within the previous five years. Participants could include tenured faculty, contract instructors, and graduate student instructors.

Recruitment

Participants for study 2 were initially recruited from the sample of individuals who had completed Study 1. At the conclusion of Study 1, participants were given the opportunity to indicate their willingness to be contacted for a follow-up study; only those who opted in were initially eligible for recruitment. Given the small sample size in Study 1 and the limited number of participants who opted to be recontacted, the recruitment contacts and procedures used in Study 1 were re-employed for Study 2. Furthermore, the wording of the recontact option in Study 1 did not require participants to actively indicate that they did not wish to be contacted in the future. As such, failure to opt in was not interpreted as a refusal of future contact.

Sample Size

Participants were statistics instructors recruited to provide evaluative feedback on the online effect size activities. Though 229 respondents used the recruitment email to visit the online survey, only 28 provided consent to data collection and only 11 provided any data beyond the demographics section.

This second study did not achieve a sample size large enough to support formal statistical analyses, but there is still a strong justification for proceeding with the study. While representative samples are necessary to make population-level generalizations or inferences, small sample studies are common for early-stage design research, particularly for when evaluations of usability are collected (Lewis, 1994). Qualitative and descriptive findings retain scholarly value even when inferential statistics are no longer appropriate given the limited sample size.

Additionally, continuing with the study demonstrates transparency and responsible research practice. Rather than contributing to the “file drawer” (Rosenthal, 1979), findings from this study should be appropriately framed as exploratory in nature, with conclusions interpreted with caution and within the constraints of the available data.

Current Academic Position

All respondents (n = 11) reported their highest degree. The majority held a PhD (n = 9), while a smaller subset held a master’s degree (n = 2). Thus, approximately 82% of the sample possessed doctoral-level credentials, indicating that participants were largely individuals with advanced research and methodological training. This distribution is consistent with the instructional contexts described in study 1, in which many instructors teaching statistics (particularly required or upper-level courses) hold terminal degrees in their field.

Years Since Degree

Participants represented a broad range of professional experience. Two respondents had earned their degree within the past five years. Three reported completing their degree more than five but less than ten years ago, and one reported completing their degree more than ten but less than fifteen years ago. Three indicated

degree completion more than fifteen but less than twenty years ago, and two reported completing their degree more than twenty years ago.

This distribution suggests that the sample included both early-career and senior instructors, enabling evaluation of the online activities from multiple professional vantage points. The presence of instructors spanning early, mid, and late career stages enhances the interpretive breadth of the findings, particularly given study 1's indication that workload distribution and teaching context may shape curricular decisions.

Teaching Rank

Teaching ranks further reflected professional diversity within the sample. Two participants identified as Associate Professors and four as Full Professors. Two identified as Lecturers, one as Contract Faculty, one selected "Other," and one identified as a PhD student or equivalent.

Although the sample size was modest, the predominance of tenured or tenure-track faculty (Associate and Full Professors; n = 6) suggests that many participants held substantial responsibility for curriculum design and instructional decision-making. At the same time, the inclusion of Lecturers, Contract Faculty, doctoral instructors, and other roles broadens the applicability of the findings across instructional positions that may differ in autonomy, workload, and curricular influence. Collectively, the sample represents a range of academic ranks and career stages, providing a broader perspective on the usability and instructional value of the online effect size activities.

Analysis

Data from study 2 were analyzed to evaluate instructor perceptions of the interactive online resources designed to support effect size estimation and interpretation. Analyses focused on both website-level design features (navigation, graphical representation,

organization) and technique-specific modules (anchor-based interpretation, standardized effect sizes, simulation-based approaches, distribution-based approaches) across seven design dimensions: content utility, purpose clarity, readability, simplicity, navigation, graphical representation, and organization.

All quantitative evaluations were collected using a consistent 10-point Likert-type scale (where 1 is *extremely dissatisfied* and 10 is *extremely satisfied*), with higher scores indicating more favorable ratings. Descriptive statistics were calculated for each design factor at both the website and module levels. These statistics provided a baseline assessment of instructors' perceptions regarding usability, clarity, pedagogical value, and overall instructional feasibility.

Website-level design factors (navigation, graphical representation, and organization) were evaluated holistically across all modules to assess general usability and structural clarity. Module-specific factors (content utility, purpose clarity, readability, and simplicity) were examined individually for each of the four effect size technique modules. This approach allowed for a nuanced understanding of how specific design elements contributed to instructor perceptions of each instructional approach.

In addition to the factor-level analysis, a global endorsement rating was collected via a single item asking instructors to indicate the likelihood of recommending the website to learners studying effect sizes. Descriptive statistics for this item provided an overall measure of the perceived instructional value of the resources, integrating considerations of usability, pedagogical clarity, and practical feasibility. The full survey can be found in Appendix B.

Results

Website Endorsement

In addition to design-factor evaluations, instructors were asked at the start of the survey to report their initial impression of the website via an open-text response field and also their overall likelihood of recommending the website to someone learning about effect sizes. This global endorsement item was measured on a 10-point scale ranging from 1 (“extremely unlikely”) to 10 (“extremely likely”).

The mean endorsement rating was 7.09 (n = 11), indicating generally favorable overall impressions of the website as an instructional resource. Although slightly lower than the highest-rated individual design features (i.e., navigation and anchor-based content utility), this score remains well above the midpoint of the scale and suggests that instructors viewed the website as recommendable to learners seeking to develop effect size estimation and interpretation skills.

Importantly, this endorsement rating reflects a holistic judgment that integrates perceptions of usability, instructional clarity, and pedagogical value. In the context of study 1, where time constraints and feasibility concerns were identified as central barriers, the positive endorsement score provides preliminary evidence that the online activities are not only well-designed on specific dimensions, but also viewed as sufficiently practical and valuable to merit recommendation within instructional settings.

Website-Level Design Features

Holistic evaluations of the website indicated strong performance in navigation, graphical representation, and organization. Navigation received the highest overall rating (M = 8.27, min = 5, max = 10, n = 11), indicating that instructors found the site easy to move through, with consistent and user-controlled access to different sections.

Graphical representation was rated favorably (M = 7.45, min = 3, max = 10, n = 11), suggesting that visual elements (including typography, layout, and visual clarity) were perceived as generally effective, though with somewhat greater variability in responses. Organization received a mean rating of 7.73 (min = 4, max = 10, n = 11), indicating that instructors largely agreed that the site structure, headings, and sequencing were logical and understandable.

Together, these findings suggest that the website-level design principles aimed at reducing navigational friction and promoting structural clarity were largely successful from the perspective of instructor users.

Qualitative feedback indicated many typographical errors were present within the lesson text which detracted from the readability and clarity of the lessons. These errors will be corrected prior to making these activities freely available online.

Content Utility Across Techniques

Content utility ratings were consistently high across all four technique modules. The anchor-based module received the highest mean rating (M = 8.36, min = 5, max = 10, n = 11), followed closely by the standardized effect size module (M = 8.09, min = 5, max = 10, n = 11). The simulation-based module was rated positively overall (M = 7.73, min = 2, max = 10, n = 11), though it exhibited the widest range of responses. The distribution-based module received a mean rating of 7.30 (min = 3, max = 10, n = 10).

These results indicate that instructors perceived the content as useful and aligned with the pedagogical purpose of teaching effect size estimation and interpretation. Notably, the comparatively high rating of the anchor-based module suggests particular

strength in activities that connect effect sizes to interpretive benchmarks, a technique that may be especially accessible within existing curricular constraints.

There is a confounding factor in that the different modules adopted slightly different design approaches. The highest rated lesson, anchor, employed a new trend called “scrollytelling” in which the interactive graphics changes in response to the user scrolling through the webpage.

Purpose Clarity

Ratings of purpose clarity were similarly favorable. The anchor-based module again received the highest mean rating (M = 8.27, min = 5, max = 10, n = 11), followed by the simulation-based module (M = 8.09, min = 4, max = 10, n = 11) and the standardized effect size module (M = 7.82, min = 4, max = 10, n = 11). The distribution-based module received a mean rating of 7.40 (min = 3, max = 10, n = 10).

These findings suggest that instructors generally perceived the instructional objectives of each module as clear and well-articulated. The relatively strong ratings for the simulation-based module are noteworthy given that simulation approaches were identified in study 1 as particularly resource-intensive and complex. Clear articulation of purpose may therefore represent one mechanism by which perceived complexity can be mitigated.

Readability

Readability ratings were positive across modules, with means ranging from 7.20 to 7.82. The simulation-based module received the highest mean readability rating (M = 7.82, min = 5, max = 10, n = 11), followed by the anchor-based module (M = 7.55, min =

5, max = 10, n = 11), the standardized effect size module (M = 7.45, min = 4, max = 10, n = 11), and the distribution-based module (M = 7.20, min = 3, max = 10, n = 10).

These results indicate that instructors found the written explanations and on-screen guidance to be clear and appropriately calibrated for a range of learners. The relatively compressed range of means across modules suggests consistency in writing quality.

Simplicity

Simplicity ratings were also favorable, with mean values ranging from 7.00 to 7.91. The anchor-based module received the highest rating (M = 7.91, min = 5, max = 10, n = 11), followed closely by the standardized effect size module (M = 7.82, min = 4, max = 10, n = 11) and the simulation-based module (M = 7.64, min = 5, max = 10, n = 11). The distribution-based module received a mean rating of 7.00 (min = 3, max = 10, n = 10).

Given that simplicity was defined in terms of ease of use, intuitive features, uncluttered layout, and device optimization, these ratings suggest that the interactive design largely succeeded in minimizing unnecessary complexity. The comparatively lower mean for the distribution-based module may reflect greater conceptual or visual density within that activity.

Summary of Design Evaluations

Across all seven design factors, instructor ratings were uniformly positive and clustered well above the midpoint of the scale. Website-level navigation emerged as a particular strength, while content utility and purpose clarity were especially strong within the anchor-based and standardized modules. Although simulation- and distribution-based modules exhibited slightly

greater variability in ratings, their mean scores remained favorable across all evaluated dimensions.

Collectively, these findings indicate that the online activities were perceived as well-designed, pedagogically purposeful, and usable across instructional contexts. In light of study 1's identification of structural and logistical barriers to expanding effect size instruction, these results provide preliminary evidence that thoughtfully designed online resources may offer a feasible and acceptable means of supporting broader integration of diverse effect size techniques.

Discussion

The purpose of study 2 was to evaluate whether carefully designed online activities could address the structural barriers to effect size instruction identified in study 1. The findings from study 1 indicated that instructors broadly value effect size instruction and possess the conceptual competence to teach it effectively, yet face persistent constraints (most notably limited instructional time and perceived technical complexity) that restrict the breadth of techniques implemented in practice. Study 2 was designed in direct response to these constraints, prioritizing usability, clarity, accessibility, and low implementation burden. The present results suggest that this design strategy has the potential to reduce these barriers.

Across all seven design dimensions, instructor evaluations were consistently positive and well above the midpoint of the scale. Website-level navigation received the strongest endorsement, indicating that instructors perceived the site as easy to move through and logically structured. This finding is particularly important in light of study 1's identification of time constraints as the most frequently cited barrier to teaching effect

size techniques. Resources that require substantial onboarding time or complex technical setup are unlikely to be adopted, regardless of conceptual merit. The high navigation and organization ratings suggest that the website minimized these friction points, thereby aligning directly with instructors' structural realities.

Content utility and purpose clarity ratings were also strong across all four technique modules. Notably, the anchor-based and standardized effect size modules received the highest content utility ratings, consistent with study 1's finding that techniques closely aligned with traditional statistical workflows tend to be more commonly taught. However, the simulation-based module (which was identified in study 1 as one of the most underutilized and barrier-prone approaches) also received favorable ratings across content utility, purpose, readability, and simplicity. Although its ratings exhibited slightly greater variability, mean scores remained high. This pattern suggests that when simulation-based techniques are embedded within clear, interactive, and self-contained online environments, some of the perceived technical and pedagogical burdens associated with teaching them may be reduced.

The endorsement data provide additional insight into the practical viability of the resource. With a mean recommendation of 7.09 out of 10 (where 1 is extremely unlikely and 10 is extremely likely), instructors indicated that they would generally be willing to recommend the website to learners studying effect sizes. While not uniformly highly rated, this level of endorsement is notable given the modest sample size and the evaluative nature of the task. Importantly, endorsement reflects a holistic judgment that integrates usability, instructional quality, and feasibility. In the context of study 1's findings, this suggests that the online activities are not merely well-designed in isolation,

but sufficiently aligned with instructional needs to be considered shareable and implementable.

Taken together, the findings of study 2 reinforce and extend those of study 1. Study 1 demonstrated that the primary challenge in expanding effect size instruction is structural rather than conceptual; instructors are willing and capable, but constrained. Study 2 provides preliminary evidence that thoughtfully developed, browser-based, interactive activities—designed for simplicity, clarity, and accessibility—may represent a viable strategy for mitigating those structural barriers. By functioning client-side, requiring no specialized software installation, and emphasizing intuitive design, the activities directly target the time and complexity concerns previously identified.

At a broader level, the convergence of findings across studies supports a pragmatic model of instructional reform. Rather than attempting to change instructor attitudes or remediate misconceptions, interventions may be more effective when they reduce logistical burden and integrate seamlessly into existing curricula. Small shifts in instructional feasibility, particularly within required statistics courses that reach thousands of students annually (as estimated in study 1), have the potential to yield substantial downstream effects on students' statistical reasoning and interpretation practices.

Limitations

Several limitations should be acknowledged. The sample size in study 2 was small, and ratings were based on instructor perceptions rather than direct measures of student learning outcomes. As such, conclusions regarding instructional effectiveness remain provisional. Future research should examine student engagement, comprehension, and

transfer of effect size reasoning following integration of these activities into live courses. Additionally, qualitative feedback may further clarify which design elements most directly influence adoption decisions.

Despite these limitations, the present findings offer encouraging evidence that design-informed, open, browser-based instructional tools can support the broader integration of diverse effect size techniques. In combination, Studies 1 and 2 suggest that advancing effect size education in the social sciences may depend less on changing what instructors believe and more on reshaping the structural conditions under which they teach.

CHAPTER FIVE:
GENERAL DISCUSSION, LIMITATIONS, AND FUTURE DIRECTIONS

General Discussion

This research investigated the instructional challenges and potential solutions for teaching effect size estimation and interpretation in undergraduate and graduate social science statistics courses. Building on study 1, which documented instructor readiness, perceived value, and structural barriers, study 2 evaluated interactive, browser-based resources designed to reduce implementation burden while enhancing conceptual clarity. Together, these studies provide a comprehensive perspective on the feasibility of advancing effect size education through design-informed interventions.

A central insight from this work is that structural constraints, rather than conceptual deficiencies or resistance, represent the primary barrier to robust effect size instruction. Instructors overwhelmingly value effect size teaching and possess the requisite conceptual competence, yet limited time, perceived technical complexity, and curricular pressures inhibit the breadth and depth of techniques they implement. These findings shift the focus of reform from persuasion or conceptual training toward reducing logistical friction and facilitating adoption.

The interactive resources developed in study 2 demonstrate that well-designed, modular, and adaptable tools can partially address these barriers. Instructor evaluations suggest that features emphasizing navigation, content utility, purpose clarity, simplicity, and readability were particularly effective in promoting usability and perceived pedagogical value. Qualitative feedback reinforced these findings, with participants frequently describing the resources as “clean”, “polished”, “inviting”, and “easy to navigate”. Several instructors highlighted the intuitive structure and accessibility of the

materials, noting that the lessons appeared appropriately pitched for undergraduate students and that the simulations and interactive visualizations supported engagement and conceptual understanding. One participant remarked that the site was “very well-done and professional” and immediately prompted the thought, “Could I use this for my stats class?”, while another commented that they “could see using one or two of [the simulations]” in their own courses. Positive reactions to dynamic visual elements, including scroll-responsive graphics (i.e., Scrollytelling) and interactive activities, further suggested that the design successfully promoted engagement and usability.

At the same time, instructors also identified several conceptual and implementation-related areas for refinement that provide useful direction for future development. Although participants generally viewed the resources positively, some questioned whether all modules maintained a sufficiently direct connection to the central goal of effect size instruction. In particular, several instructors noted that simulations and probability-focused activities, while engaging and pedagogically valuable, did not always clearly demonstrate how they advanced students’ understanding of effect sizes specifically. The anchor-based module was frequently viewed as having the strongest alignment with the website’s stated instructional objectives, whereas some participants suggested that the distribution and simulation sections could benefit from a more explicit connection between abstract statistical ideas and effect size interpretation.

Participants also provided constructive recommendations for improving instructional coherence and practical relevance. Suggestions included integrating more real-world research examples, presenting field-specific benchmarks rather than relying solely on Cohen’s conventional thresholds, and clarifying the pragmatic purpose of certain

conceptual components. Several instructors recommended reorganizing portions of the site, such as placing statistical refresher materials earlier in the sequence, and replacing or simplifying cognitively demanding examples like the Monty Hall simulation with more intuitive alternatives. Others emphasized the importance of refining the written content through improved grammar, typography, visual hierarchy, and more engaging graphical design elements to better sustain student attention and reduce cognitive load.

Importantly, these critiques were generally framed within otherwise favorable evaluations of the website's instructional potential. Even instructors who expressed reservations about scope or alignment frequently described the simulations and interactive activities as engaging, useful, and adaptable for classroom use. Collectively, the feedback suggests that the primary challenge moving forward is not the perceived value of interactive instructional approaches themselves, but rather ensuring tighter conceptual integration, clearer pedagogical framing, and stronger alignment between individual activities and the overarching learning objectives surrounding effect size reasoning and interpretation. At the same time, these findings reinforce a broader implication emerging across both studies: instructors are generally receptive to effect size instruction and innovative teaching approaches when those approaches are practical, accessible, and feasible within existing instructional constraints.

Taken together, the studies suggest a pragmatic model for instructional reform. Rather than attempting to change instructor beliefs (which are already aligned with best practices), interventions may be more effective when they reshape the structural and technological conditions under which teaching occurs. Small improvements in feasibility, particularly within required courses that reach large numbers of students, have the

potential for meaningful downstream impact on students' statistical reasoning and interpretation skills. Central to this broader model of reform is the development of instructional resources that are not only pedagogically effective, but also flexible, accessible, and sustainable within diverse teaching contexts.

Another notable insight is the role of modular, open, and scalable resources in promoting instructional flexibility. By leveraging widely available technologies (D3, R, standard web platforms) and emphasizing client-side interactivity, the resources provide instructors with low-barrier, adaptable tools. This open design philosophy aligns with principles of sustainability and collaborative improvement, allowing instructors and researchers to modify, extend, or redistribute resources as needed.

Limitations

Several limitations temper the conclusions drawn from this research. First, the sample size in study 2 was modest and heavily skewed toward instructors with advanced methodological training and doctoral-level education. While this aligns with the population most likely to teach effect size in formal settings, it limits the generalizability of the findings to instructors with less statistical experience or in non-research-intensive institutions. Second, the evaluation relied primarily on self-report measures, which are susceptible to social desirability bias. Participants may have been inclined to provide more favorable evaluations of the website's usability, relevance, or perceived effectiveness than they genuinely held, particularly given their awareness of the study's objectives. As a result, the findings may overestimate participants' true perceptions and experiences with the resource. Third, the assessment was formative and short-term, capturing initial impressions and perceived utility rather than sustained

use over multiple semesters or long-term adoption trends. Fourth, the scope of effect size techniques addressed in the interactive resources was intentionally limited; certain underrepresented measures, such as health economic effect sizes or context-specific indices, were introduced only conceptually, potentially underestimating the challenges associated with teaching these approaches. Finally, the effectiveness of the resources presupposes a baseline level of technological and statistical literacy among instructors, which may vary across institutional and regional contexts, and could influence adoption and implementation success. Collectively, these limitations suggest caution in extrapolating the results broadly and highlight the need for more extensive, diverse, and longitudinal evaluation.

Future Research

Future research should extend these findings by directly assessing student learning outcomes in both controlled and naturalistic classroom contexts. Understanding whether interactive modules improve interpretive reasoning, conceptual understanding, and the ability to apply effect size knowledge across contexts is critical. Additionally, investigations should broaden instructor samples to include diverse institutional types, teaching experience levels, and international settings to evaluate generalizability and contextual constraints. Longitudinal studies examining sustained adoption, integration into full curricula, and the downstream impact on student statistical literacy will provide stronger evidence of practical feasibility. Finally, iterative refinement of interactive modules, informed by instructor and student feedback, can enhance design features such as scaffolding, visual clarity, and cognitive load management, while exploring underrepresented effect size techniques to further support broad adoption.

Concluding Remarks

This research underscores that advancing effect size instruction in the social sciences requires more than instructor willingness; it demands structural and technological support that makes adoption feasible, scalable, and pedagogically sound. Interactive, browser-based resources (designed with usability, clarity, and low instructional burden in mind) offer a promising pathway toward broader integration of diverse effect size techniques. By addressing structural barriers rather than attempting to change conceptual attitudes, this work contributes both to statistics education reform and to the broader understanding of how design-informed interventions can enhance instructional feasibility.

REFERENCES

Adkins, M. C. (2024, April 30). Mind the gap: Assessing current pedagogical approaches to teaching effect size estimation and interpretation in university level psychology statistics classrooms. OSF. https://osf.io/vbj7y/

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology, 74, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes: The handling of quantitative results in published *Health Education & Behavior* research. *Health Education & Behavior, 43*(5), 518–527. https://doi.org/10.1177/1090198116669521

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics, 108*, 355–361. https://doi.org/10.1111/j.1949-8594.2008.tb17850.x

Beribisky, N., Alter, U., & Cribbie, R. A. (2025). A multi-faceted mess: A review of statistical power analysis in psychology journal articles. *Meta-Psychology, 9*. https://doi.org/10.15626/MP.2020.2643

Bostock, M. (2021). D3.js (Version 7.0) [Computer software]. D3. d3js.org

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G. H., Velleman, P., Witmer, J., & Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. American Statistical Association.

Carver, R. H., & Stephens, M. (2014). It is time to include data management in introductory statistics. In Proceedings of the Ninth International Conference on Teaching Statistics. International Association for Statistical Education.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. Technology Innovations in Statistics Education, 1(1). https://doi.org/10.5070/T511000026

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of educational research*, *90*(4), 499-541. https://doi.org/10.3102/003465432093354

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? Technology Innovations in Statistics Education, 1(1). https://doi.org/10.5070/T511000028

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology, 65(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1992). Statistical power analysis. Current Directions in Psychological Science, 1(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Collins, E., & Watt, R. (2021). Using and understanding power in psychological research: A survey study. Collabra: Psychology, 7(1), Article 28250. https://doi.org/10.1525/collabra.28250

Cook, J. A., Hislop, J., Adewuyi, T. E., Harrild, K., Altman, D. G., Ramsay, C. R., Campbell, M. K., et al. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. Health Technology Assessment, 18(28), 1–550. https://doi.org/10.3310/hta18280

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. Zeitschrift für Psychologie / Journal of Psychology, 217(1), 15–26. https://doi.org/10.1027/0044-3409.217.1.15

Cumming, G. (2014). The new statistics: Why and how. Psychological Science, 25(1), 7–29. https://doi.org/10.1177/0956797613504966

delMas, R. C., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. Statistics Education Research Journal, 6(2), 28–58. https://doi.org/10.52041/serj.v6i2.483

Doo, M. Y., Bonk, C., & Heo, H. (2020). A meta-analysis of scaffolding effects in online learning in higher education. International Review of Research in Open and Distributed Learning, 21(3), 60-80. https://doi.org/10.19173/irrodl.v21i3.4638

Edelsbrunner, P. A., & Thurn, C. M. (2024). Improving the utility of non-significant results for educational research: A review and recommendations. Educational Research Review, 42, Article 100590. https://doi.org/10.1016/j.edurev.2023.100590

Farmus, L., Beribisky, N., Martinez Gutierrez, N., Alter, U., Panzarella, E., & Cribbie, R. A. (2023). Effect size reporting and interpretation in social personality research. Current Psychology, 42, 15752–15762. https://doi.org/10.1007/s12144-021-02621-7

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K–12 curriculum framework. American Statistical Association.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. Journal of Experimental Psychology: General, 141(1), 2–18. https://doi.org/10.1037/a0024338

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156–168. https://doi.org/10.1177/2515245919847202

Garett, R., Chiu, J., Zhang, L., & Young, S. D. (2016). A literature review: Website design and user engagement. Online Journal of Communication and Media Technologies, 6(3), 1–14. https://doi.org/10.29333/ojcmt/2556

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. Statistics Education Research Journal, 4(1), 92–99. https://doi.org/10.52041/serj.v4i1.527

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. International Statistical Review, 75(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning: Connecting research and teaching practice. Springer.

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. Teaching Statistics, 31(3), 72–77. https://doi.org/10.1111/j.1467-9639.2009.00363.x

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. Journal for Research in Mathematics Education, 19(1), 44–63. https://doi.org/10.5951/jresematheduc.19.1.0044

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., & Murphy, T. J. (2015). Identifying statistical concepts associated with high and low levels of self-efficacy to teach statistics in middle grades. Journal of Statistics Education, 23(1). https://doi.org/10.1080/10691898.2015.11889724

Huberty, C. J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62(2), 227–240. https://doi.org/10.1177/0013164402062002002

International Organization for Standardization. (2019). Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210:2019). https://www.iso.org/standard/77520.html

Kelley, K., & Preacher, K. J. (2012). On effect size. Psychological Methods, 17(2), 137–152. https://doi.org/10.1037/a0028086

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. American Psychologist, 56(1), 16–26. https://doi.org/10.1037/0003-066X.56.1.16

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. BMC Research Notes, 8, Article 84. https://doi.org/10.1186/s13104-015-1020-4

Lakens, D. (2022). Sample size justification. Collabra: Psychology, 8(1), Article 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. Advances in Methods and Practices in Psychological Science, 4(1). https://doi.org/10.1177/2515245920951503

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. Human Factors, 36(2), 368–378. https://doi.org/10.1177/001872089403600215

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. The American Statistician, 54(3), 196–206. https://doi.org/10.1080/00031305.2000.10474545

Loyens, S. M. M., & Gijbels, D. (2008). Understanding the effects of constructivist learning environments: Introducing a multi-directional approach. Instructional Science, 36, 351–357. https://doi.org/10.1007/s11251-008-9059-4

Lu, Y. (2023). Web-based applets for facilitating simulations and generating randomized datasets for teaching statistics. Journal of Statistics and Data Science Education, 31(3), 264–272. https://doi.org/10.1080/26939169.2022.2146614

Lysakowski, R. S., & Walberg, H. J. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. American Educational Research Journal, 19(4), 559–572. https://doi.org/10.3102/00028312019004559

McNamara, A. (2019). Key attributes of a modern statistical computing tool. The American Statistician, 73(4), 375–384. https://doi.org/10.1080/00031305.2018.1482784

Morris, P. H. (2020). Misunderstandings and omissions in textbook accounts of effect sizes. *British Journal of Psychology*, *111*(2), 395–410. https://doi.org/10.1111/bjop.12401

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. Psychological Methods, 23(2), 208–225. https://doi.org/10.1037/met0000126

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. Educational Psychology Review, 25(2), 157–209. https://doi.org/10.1007/s10648-013-9218-2

Piaget, Jean. 1953. *The Origin of Intelligence in the Child*. London: Routledge & Kegan Paul.

Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. Computers & Education, 58(1), 136–153. https://doi.org/10.1016/j.compedu.2011.07.017

Schmidt, K., Edelsbrunner, P. A., Rosman, T., Cramer, C., & Merk, S. (2023). When perceived informativity is not enough: How teachers perceive and interpret statistical results of educational research. Teaching and Teacher Education, 130, Article 104134. https://doi.org/10.1016/j.tate.2023.104134

Schneider, F., van Osch, L., & de Vries, H. (2012). Identifying factors for optimal development of health-related websites: A Delphi study among experts and potential future users. Journal of Medical Internet Research, 14(1), Article e18. https://doi.org/10.2196/jmir.1863

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. https://doi.org/10.1080/09500693.2011.605182

Sokolowski, A., Li, Y., & Willson, V. (2015). The effects of using exploratory computerized environments in grades 1 to 8 mathematics: A meta-analysis of

research. International Journal of STEM Education, 2, Article 8. https://doi.org/10.1186/s40594-015-0022-z

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. Journal of Educational Psychology, 102(4), 989–1004. https://doi.org/10.1037/a0019507

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. The American Statistician, 69(4), 362–370. https://doi.org/10.1080/00031305.2015.1081619

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. Journal of Statistics Education, 20(2), 1–24. https://doi.org/10.1080/10691898.2012.11889641

Vlachogianni, P., & Tselios, N. (2022). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. Journal of Research on Technology in Education, 54(3), 392–409. https://doi.org/10.1080/15391523.2020.1867938

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zavez, K., & Harel, O. (2025). Teaching statistical concepts using computing tools: A review of the literature. Journal of Statistics and Data Science Education, 1–12. https://doi.org/10.1080/26939169.2024.2445541

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. Statistics Education Research Journal, 7(2), 40–58. https://doi.org/10.52041/serj.v7i2.469

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. Journal of Statistics Education, 16(2). https://doi.org/10.1080/10691898.2008.11889566

APPENDIX A: STUDY 1 QUESTIONNAIRE

Educational Background and Experience

1. **What is the highest post-secondary degree you have obtained?**

□ Master's
□ PhD

2. **How many years ago was your highest degree completed? (Optional)**

□ Less than 5 years ago
□ More than 5 years ago, but less than 10 years ago
□ More than 10 years ago, but less than 15 years ago
□ More than 15 years ago, but less than 20 years ago
□ More than 20 years ago

3. **Which of the following titles best applies to your current position? (Optional)**

□ Research or Teaching Employee without a PhD
□ PhD student or equivalent
□ Postdoctoral researcher
□ Lecturer
□ Assistant Professor
□ Associate Professor
□ Full Professor
□ Contract Faculty
□ Other
□ Prefer not to say

Logic: If "Other" selected:

Please specify your current position.

[] Open text

4. **Which sub-field of psychology do you identify with? E.g., sports psychology, health psychology, neuropsychology. Select one from the list or type in your response. (Optional)**

□ Addiction Psychology
□ Adult Development and Aging
□ Black Psychology
□ Brain and Cognitive Sciences
□ Clinical Psychology
□ Clinical Neuropsychology
□ Community Psychology

□ Counselling Psychology
□ Criminal Justice Psychology
□ Developmental Psychology
□ Educational and School Psychology
□ Environmental Psychology
□ Extremism and Terrorism
□ Family Psychology
□ Health Psychology and Behavioural Medicine
□ History and Philosophy
□ Indigenous Peoples' Psychology
□ Industrial/Organizational Psychology
□ International and Cross-Cultural Psychology
□ Psychologists in Hospitals and Health Centres
□ Psychology in the Military
□ Psychologists and Retirement
□ Psychopharmacology
□ Quantitative Methods
□ Quantitative Electrophysiology
□ Rural and Northern Psychology
□ Sexual Orientation and Gender Identity
□ Social and Personality
□ Sport and Exercise Psychology
□ Teaching of Psychology
□ Traumatic Stress
□ Section on Women and Psychology (SWAP)

□ Other (please specify): _____

Course Levels Taught

Instructions: "This section is quite **important**. Please answer it carefully because the structure of the remaining questions is based upon your answers. Once you click **NEXT**, you will be unable to change your responses. For each level of course you taught, there will be a series of questions about those courses specifically. Textual prompts and colours will be used to remind you in each section that the questions pertain to either your undergraduate or your graduate level courses."

5. **Select all levels of courses you have taught.**

□ Undergraduate
□ Graduate

6. **How comfortable are you teaching students to estimate and interpret effect sizes within university-level statistics courses?**

1 = Not at all comfortable
7 = Extremely comfortable

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

7. **How important is teaching students how to interpret and estimate effect sizes for their statistical education?**

□ Not at all important
□ Not very important
□ Somewhat important
□ Important
□ Very important

Effect Size Knowledge

Instructions: "Your next task is to answer 'True' or 'False' to each of the following statements. There may or may not be an equal number of true and false responses. Please select 'I don't know' instead of making a random guess."

8. **Effect sizes express the magnitude of the influence one variable has on another.**
□ True
□ False
□ I don't know

9. **A larger sample size will result in a larger (stronger) effect size.**
□ True
□ False
□ I don't know

10. **If you are doing high-quality research, your aim is to find the largest effect sizes possible.**
□ True
□ False
□ I don't know

11. **A statistically significant p-value corresponds to a medium or large effect size.**
□ True
□ False
□ I don't know

12. **A small effect size indicates that the null hypothesis should fail to be rejected.**

□ True
□ False
□ I don't know

Undergraduate-Level Effect Size Techniques

Logic: Participants who indicated that they taught undergraduate statistics courses completed this section.

Instructions: “For each of the following techniques, please indicate the statement that best reflects your use of the technique in undergraduate-level courses.”

	I have used this technique in the classes I teach	I am familiar with this technique, but I do not use it in the classes I teach	I am not familiar with this technique
13. Anchor-based interpretation	□	□	□
14. Distribution-based interpretation	□	□	□
15. Health economic (cost-benefit analysis)	□	□	□
16. Opinion-seeking (expert opinion)	□	□	□
17. Pilot-study approach	□	□	□
18. Review of evidence base (meta-analysis or literature review)	□	□	□
19. Standardized effect sizes (e.g., Cohen's d)	□	□	□
20. Simulation-based effect sizes	□	□	□

21. **Is there any other technique missing from the list above which should have been included?**

□ Yes
□ No

Logic: If "Yes" was selected for Q21, participants completed the following items:

21a. **What is a short name for this technique?**

[] Open text

21b. **Could you describe the missing technique?**

[] Open text

21c. **Please indicate your familiarity with this technique using the same three response options presented above.**

□ I have used this technique in the classes I teach

□ I am familiar with this technique, but I do not use it in the classes I teach

□ I am not familiar with this technique

<u>Undergraduate-Level Instruction: Introductory and Advanced Courses</u>

Logic: Participants who indicated that they taught undergraduate statistics courses were asked the following question:

22. **Which of the following best describes the types of undergraduate statistics courses you teach?**

□ Introductory only
□ Advanced only
□ Introductory and advanced

*Logic: Participants who selected **Introductory and Advanced** completed the following section, all other single options skip the participant to the next section.* Participants were shown only those techniques that they had previously indicated using in their undergraduate teaching.

Instructions: "For each technique that you indicated using in your undergraduate teaching, please identify whether you teach the technique in introductory courses, advanced courses, or both. If you teach a technique in both types of courses, select both response options."

23. **Anchor-based interpretation**

□ Introductory
□ Advanced

24. **Distribution-based interpretation**

□ Introductory
□ Advanced

25. **Health economic (cost-benefit analysis**

□ Introductory
□ Advanced

26. **Opinion-seeking (expert opinion)**

□ Introductory
□ Advanced

27. **Pilot-study approach**

□ Introductory
□ Advanced

28. **Review of evidence base (meta-analysis or literature review)**

□ Introductory
□ Advanced

29. **Standardized effect sizes (e.g., Cohen's d)**

□ Introductory
□ Advanced

30. **Simulation-based effect sizes**

□ Introductory
□ Advanced

31. **Other technique identified by participant (if applicable).**

□ Introductory
□ Advanced

Barriers to Teaching Effect Size Techniques (Undergraduate Level)

Logic: Participants who indicated that they were familiar with but did not use at least one effect size technique in their undergraduate teaching were asked to complete the following section.

Instructions. "For each technique you identified as familiar but not currently used in your undergraduate teaching, please indicate the reasons why you do not teach the technique. Select all that apply."

32. Anchor-based interpretation

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique

□ Prefer not to say
□ Other (please specify)

33. Distribution-based interpretation

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

34. Health economic (cost-benefit analysis)

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

35. Opinion-seeking (expert opinion)

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

36. Pilot-study approach

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

37. Review of evidence base (meta-analysis or literature review)

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

38. Standardized effect sizes (e.g., Cohen's d)

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

39. Simulation-based effect sizes

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

40. Additional technique identified by participant (if applicable)

□ This technique is too technical/complex to teach at this level
□ I am not familiar enough with this technique to teach it at this level
□ There is not enough class time to teach this technique
□ Prefer not to say
□ Other (please specify)

Undergraduate Course Attributes

Logic: Participants who reported teaching undergraduate statistics courses completed the following items regarding the characteristics of those courses over the past five years.

41**. How many different statistics courses did you teach during the last 5 years (e.g., different course codes)?**

___ numeric response ___

42. **How many of those courses were required introductory statistics classes?**

___ numeric response ___

43. **What was the smallest class size (approximately)?**

___ numeric response ___

44. **What was the largest class size (approximately)?**

___ numeric response ___

45. **On average, what percentage of a student's grade was based on interpreting or estimating effect sizes (e.g., exam questions, assignments)?**

0% – 100% (continuous scale)

46. **Select all methods that students used to compute effect sizes in your courses:**

□ ESCI
□ Excel
□ G*Power
□ Hand calculations
□ JAMOVI
□ JASP
□ MATLAB
□ Mplus
□ Online calculator
□ Python
□ R
□ SAS
□ SPSS
□ Stata
□ None (students did not compute effect sizes)
□ Other (please specify): _____

Logic: For users who indicated that they taught graduate level courses, the following sections from earlier in the survey were repeated with language altered to highlight that the questions pertained to their graduate level teaching. The same logic and series of questions were repeated where appropriate given the user's responses.

- **Undergraduate-Level Effect Size Techniques**
- **Undergraduate-Level Instruction: Introductory and Advanced Courses**
- **Barriers to Teaching Effect Size Techniques (Undergraduate Level)**
- **Undergraduate Course Attributes**

APPENDIX B: STUDY 2 QUESTIONNAIRE

Educational Background and Experience

Instructions: “An instructor's educational background and experience are important information. This short section will be about your education and current academic position.”

1. **What is the highest post-secondary degree you have obtained?**
□ Master's
□ PhD

2. **How many years ago was your highest degree completed? (Optional)**
□ Less than 5 years ago
□ More than 5 years ago, but less than 10 years ago
□ More than 10 years ago, but less than 15 years ago
□ More than 15 years ago, but less than 20 years ago
□ More than 20 years ago

3. **Which of the following titles best applies to your current position? (Optional)**
□ Research or Teaching Employee without a PhD
□ PhD student or equivalent
□ Postdoctoral researcher
□ Lecturer
□ Assistant Professor
□ Associate Professor
□ Full Professor
□ Contract Faculty
□ Other
□ Prefer not to say

Logic: If "Other" selected:

3a. Please specify your current position.

[] Open text

What was your initial impression of the website?

4. **What was your initial impression of the website?**

[] Open text

Website Recommendation

5. **How likely are you to recommend this website to someone learning about effect sizes?**

1 = Extremely unlikely

10 = Extremely likely

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Overall Site

Instructions: “This brief section of the survey will ask about design factors which are applicable sitewide. Please think about your overall experience of the website when answering the following questions.”

6. **Overall Site: Navigation**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

7. **Overall Site: Graphical Representation**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

8. **Overall Site: Organization**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Effect Size Techniques

Logic: There were 4 blocks of questions pertaining to each of the four technique sections of the website. The order these blocks appeared to participants was randomized by the survey software. Below is an example of one of those sections.

Instructions: “This brief section of the survey will ask about features as they apply to the lessons within the ‘simulation based’ section of the website. Please think about only the content in this section when answering these questions.”

9. **Simulation Based: Content Utility**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

10. **Simulation Based: Purpose**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

11. **Simulation Based: Simplicity**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

12. **Simulation Based: Readability**

1 = Extremely dissatisfied

10 = Extremely satisfied

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Time Spent Using the Website/Activities

13. **Approximately how much time (in minutes) did you spend using the website/activities?**

____numeric_____

min: 10, max: 1000

Feedback

14. **Do you have any additional comments or suggestions which could help improve this website? (Optional)**

[] Open text