Spring 2020

Technology Effectiveness in Teaching Statistics: Best-Evidence Meta-Analysis

Zita Tiamuh

TECHNOLOGY EFFECTIVENESS IN TEACHING STATISTICS: BEST-EVIDENCE META-ANALYSIS

by

Zita Tiamuh

A Dissertation
Submitted to the Graduate School,

and the School of Education
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:

Dr. Kyna Shelley, Committee Chair
Dr. Peter Paprzycki
Dr. Richard Mohn
Dr. Thomas Lipscomb
Dr. Thomas V. O'Brien

Dr. Kyna Shelley	Dr. Sandra Nichols	Dr. Karen S. Coats
Committee Chair	Director of School	Dean of the Graduate School

May 2020

Published by the Graduate School

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

ABSTRACT

Evidence about technology effectiveness in supporting post-secondary students' learning of introductory statistics concepts is inconclusive. Lacking in current investigations are considerations of the synergies between technology, content, and pedagogy that influence learning outcomes in statistics education. The current study used meta-analytic procedures to address the gap between theory and practice related to the best evidence of effective instructional practices in technology-enhanced introductory statistics classrooms. A conceptual framework based on the ADDIE model, TPACK, and constructivism guided the investigation of substantive study characteristics related to instructional design.

Findings were based on 32 studies published between 1998-2018 that used quasi-experimental or experimental research designs and measured statistics achievement. Hedges' *g* effect sizes were computed for each study used in the meta-analysis. Random-effects analysis revealed a small average effect of 0.23 favoring technology use over no technology control conditions. Mixed-effects results revealed instructional design characteristics that were significant moderators, favoring technology use. Concerning the learning context, significant effects were found among studies with undergraduate student samples (0.45), discipline-specific courses (0.31), and studies with learning goals associated with statistical literacy, thinking, or reasoning (0.42) and learning statistical skills/concepts (0.28). Regarding content, design, and duration, significant effects were found among studies covering descriptive or null hypothesis testing (0.74), that used technology designed by the instructor (0.30) and for a semester or longer (0.25). Significant effects for instruction implementation included the use of various learning

tasks (0.33), students' cooperative, collaborative, or collective engagement (0.38), use of scaffolding (0.36), and the use of technology with multiple functions for covering concepts (0.42). Concerning assessment, significant effects were found for studies using multiple formative assessment measures (0.34) and those using non-authentic assessments (0.28).

Non-significant results were found for report and methodological characteristics, except for studies whose description of the instructional design process was somewhat replicable (0.36). Sensitivity analyses did not indicate publication bias. However, interpretation of meta-analysis findings should be made with considerations that findings are based mostly on studies with quality ratings of unclear risk of bias (63%). Findings are discussed in light of the literature. Implications and recommendations for future research are provided.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and thanks to the individuals who offered unending guidance, support, and encouragement that resulted in the successful completion of this academic endeavor. I would like to first give acknowledgement and praise to my Lord and savior whose guidance through the Holy Spirit was the root and cause of my persistence throughout this journey. Next, without a second thought, I give a big thank you to my advisor and chair, Dr. Kyna Shelley. With great wisdom, grace, guidance, and endless encouragement, you saw me through my most challenging times throughout this endeavor and never failed to remind me of my potential to accomplish what I set out to do. Additionally, I thank my committee members, Dr. Paprzicki, Dr. Lipscomb, Dr. Mohn, and Dr. O'Brien for their knowledge, expertise, and support in meeting the various milestones throughout this process.

I would be remiss if I did not acknowledge the contributions of my family and dear friends. Despite the inability of some to directly relate to the intellectual, emotional, and physical fortitude required to persist through this unique dissertation endeavor, they all stood by me with love and uplifting words.

To my siblings and extended family, thank you for believing in my pursuit and for your support from the beginning until the end. To Marsha Johnson, thank you for being an open ear, a mother-figure to turn to, and a spiritual well of encouragement. To the Ambeau family, your endless support both near and afar is forever cherished and remembered. To my dear friends, thank you for being my cheerleaders on the side and for always believing in me. To Mihili Edirisooriya, I am forever grateful for your valuable contribution in supporting my persistence through this dissertation experience. Finally,

last but certainly not least, I give thanks to my RESA program colleagues. I thank you for your endless help, support, and encouragement. I could not ask for a better group of individuals to share this experience with than with you.

DEDICATION

This dissertation is dedicated in memory of my beloved ones whose lives on earth were too short-lived. These include my late father, Dr. Aloysius Tiamuh, mother, Florence Tiamuh, brother, Brice Tiamuh, sister, Olivia Tiamuh, and my precious baby niece Florence Tazifor. May their beautiful souls continue to rest in peace. It is through their memories that I smile on this day and during days where the challenge of this endeavor was heart-felt. I never had the privilege of letting my father know that as a young child in secondary school, it was because of the grit that he modeled in accomplishing his Ph.D. in Criminology that inspired me and ignited in me the desire to someday achieve the same. As both my parents demonstrated a strong value for education through their graduate degree accomplishments, they did so with the intent to transfer and apply the acquired knowledge to make a difference in others' lives. As both were educators in their own respect, it was through their own lives and the lives of those they positively impacted that has fueled in me my own passion to educate and equip others so that they may live productive and fulfilled lives. This dedication honors these family members because of the piece of my heart that each captured, their unique influences in shaping who I am today, and memories that were shared both directly and indirectly. Although they are not physically present to share in the celebration and joy of this moment, I know that they are rejoicing in spirit because of my accomplishment in receiving my Doctor of Philosophy degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I - INTRODUCTION

Research investigations concerned with assessing the role, impact, or effectiveness of technology use on a variety of education-related outcomes have been prominent in education research including statistics education research (Eichler & Zapata-Cardona, 2016; Garfield & Ben-Zvi, 2007; Mcgrath, 2014; Tishkovskaya & Lancaster, 2012). This has been amidst a general awareness of the affordances provided by educational technologies for supporting cognitive, affective, and behavioral learning outcomes (Chance, Ben-Zvi, Garfield, & Medina, 2007; Kennewell, 2001; Lowerison, Sclater, Schmid, & Abrami, 2006; Tishkovskaya & Lancaster, 2012; Xu, Zhang, Su, Cui, & Qi, 2014). In statistics education, the importance of technology has been emphasized in reform initiatives as it supports conceptual-based learning, collaboration, student engagement, data exploration, manipulation, visualization, action-oriented, and task-based individualized learning (Lloyd & Robertson, 2012; GAISE College Report ASA Revision Committee, 2016; Xu et al., 2014). Furthermore, these technologies include hardware and software tools associated with data analysis, computation, graphic and visualization, drill and practice, tutorials, multi-media learning, simulation, Internet, real data, communication technologies, and learning management systems (Garfield & Ben-Zvi, 2004; Lajoie, 1997).

The goal of technology effectiveness research is to gain an understanding if and how technology use gives rise to student learning (Lowyck, 2014; Schrum et al., 2007; Spector et al., 2014). These studies are conducted through primary research and meta-analysis research methods. Experts have heeded the need to improve the quality of studies, calling for research that clearly highlights the technological affordances

(potential uses/capabilities) associated with pedagogical practices and subject matter content. It is argued that this perspective, rather than a focus on technological features and characteristics alone is necessary for evaluating technology effectiveness (Harris, Mishra, & Koehler, 2009; Mishra, Koehler, & Bragg, 2006; Roblyer, 2005; Schrum et al., 2007; Thompson et al., 2008).

Among the various disciplinary areas covered in the research on technology effectiveness, statistics education has been regarded as one that is significantly impacted by technology innovations, attributed to the changes that have occurred in how the subject is taught, as well as the type of content matter covered (Chance, Ben-Zvi, et al., 2007; Tishkovskaya & Lancaster, 2012; Xu et al., 2014). Technological tools have allowed ease and automation of complex calculations, with less emphasis placed on mathematical computations and more focus on teaching and learning difficult, but fundamental concepts such as probability, variation, and randomness (Chance, Ben-zvi, et al., 2007; DelMas et al., 2007; Lowerison et al., 2006). Furthermore, the push for technology integration in the classroom has been followed by the assessment of its effectiveness on statistical learning outcomes (Cobb, 1992; GAISE College Report ASA Revision Committee, 2016; Hassad, 2014). This has resulted in an increase in primary research assessing the effectiveness of pedagogical practices and reform initiatives related to technology use in statistics education classrooms (Chance, Ben-Zvi, et al., 2007; Tishkovskaya & Lancaster, 2012).

Primary studies on technology effectiveness in post-secondary statistics education have been conducted in the context of teaching, focusing on various modes of classroom instructional delivery, using different types of technological tools to support learning

(such as graphing calculators, statistical software, tutorials, applets, clickers, etc.), and have compared student achievement outcomes among students using technology and those not using technology (Chance, Ben-zvi, et al., 2007; Lachem, 2014; Lloyd & Robertson, 2012; Peterson, 2016; Phillips & Phillips, 2016; Schwier & Seaton, 2013). Conclusions made about student achievement have often been based on learning outcomes associated with course grades, exam grades, projects/assignments, course evaluations, and students' self-reported perceptions of learning (e.g. affective outcomes).

Though many primary studies have concluded advantages in using technology compared to not using technology, other studies have reported no difference or negative effects on students' statistical achievement (Larwin & Larwin, 2011; Schenker, 2007). This has also come with the recognition that although the idea of adopting student-centered and active learning approaches is well-accepted, the actual integration of these practices can be challenging to educators (Roseth, Garfield, & Ben-Zvi, 2008; Tishkovskaya & Lancaster, 2012).

Given these concerns, education researchers and statistics education researchers have used meta-analysis techniques to investigate the overall effectiveness of using technologies to support student learning and to identify the various features of technology use that influence its effects on learning (Hsu, 2003; Schenker, 2007; Sosa et al., 2011; Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011).) Furthermore, meta-analysis is a useful approach for quantitatively addressing research questions about a phenomenon when a large number of primary studies exist that investigate the same topic (Cooper, 2017; Cooper & Hedges, 1994), and when inconsistent results are reported in the literature (Cooper, 2017; Tamim et al., 2011).

Generally, meta-analyses on technology effectiveness in general education and statistics education have reported small to medium positive effects of technology use on student achievement compared to not using technology (Archer et al., 2014; Hsu, 2003; Larwin & Larwin, 2011; Schenker, 2007; Schmid et al., 2014; Sosa et al., 2011; Tamim et al., 2011). Using a variety of quantitative techniques adopted from traditional analysis methods (e.g. meta-regression, ANOVA, hierarchical linear model, etc.), meta-analyses in statistics education have explored the influence of a variety of substantive study and methodological characteristics (variables) as potential moderators of the effect of technology use on student achievement. Commonly-examined study characteristics have included disciplinary field, course type, student academic level, course level, type of technology, technology feature, technology function, duration of technology use, learner control, and mode of instruction, (Hsu, 2003; Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011). Meanwhile, examples of methodological characteristics examined have included publication year, publication source, randomization of participants, and instructor bias (Hsu, 2003; Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011). Furthermore, these studies have examined the effects of technology use on cognitive outcomes (e.g. student achievement – course grades, exam grade, quiz grade, etc.) (Larwin & Larwin, 2011) as well as affective (e.g. anxiety, attitude) (Schenker, 2007) measured in primary studies.

Concerns are raised about the approaches used to select primary studies that differ in quality, as well as those that differ in their units of analysis, research designs, and statistical analysis methods employed (Kock, 2009; Rosenthal & DiMatteo, 2001). These have been referred to as comparing “garbage-in garbage-out”, and “oranges and apples”,

respectively (Cooper & Hedges, 1994; Kock, 2009; Rosenthal & DiMatteo, 2001). These differences in approaches lead to biased and conflicting conclusions (Borenstein, Hedges, Higgins, & Rothstein, 2017; Cooper, 2017). According to Cooper, Hedges, and Valente (2009) the variation in the level of rigor used to conduct primary studies that are contained in research syntheses has an impact on the conclusions made by the meta-analyst. Thus, the judgment of study quality is necessary to assess the validity of conclusions made in primary studies. Furthermore, Cooper, Hedges, and Valente (2009) define study quality as "the fit between a study's goals and the study's design and implementation characteristics." (p. 138)

Generally recommended and accepted, is the use of a broad and exhaustive criteria to select primary studies and code variables for meta-analysis (Glass, McGraw, & Smith, 1981; Stock, 1994). However, this approach has been criticized for resulting in meta-analyses that examine large numbers of primary studies that include those of low quality, which potentially weakens the analyses (Hunter & Schmidt, 2004; Slavin, 1995). It has been argued that the quality of studies selected should be of high consideration (Hunter & Schmidt, 2004; Slavin, 1995) and the coding of items should be based on conceptual or theoretical justifications (Card, 2012; Hunter & Schmidt, 2004; Slavin, 1995). Counterarguments have emphasized that though a broad selection criterion may lead to the inclusion of studies with weak methodological quality, these variables may explain other variations due to differences in methodological characteristics (Cooper, 2017; Cooper & Hedges, 1994). Furthermore, concerning the coding of studies, the goal is to focus on features and characteristics that are most relevant and are based on knowledge of the area under study (Card, 2012; Glass et al., 1981; Stock, 1994).

Problem statement

The benefits of using technologies to support post-secondary teaching and student achievement compared to not using them have been well documented in general education and statistics educational literature through primary and meta-analysis studies (Bernard et al., 2009; Sosa et al., 2011). However, in spite of this, current research assessing technology effectiveness point to three main concerns associated with: 1) the accurate assessment of the effectiveness of technology use on student achievement; 2) the need to enhance methodological approaches in meta-analysis research; and 3) the usefulness of findings for most effectively integrating technology in statistics classrooms to support student learning (Chance, Ben-zvi, Garfield, & Medina, 2007; Pearl et al., 2012).

Due to the complex nature of the classroom environment, accurately assessing the effectiveness of technology integration becomes a challenge (Morrison & Ross, Steven, 2014; Robinson et al., 2009). This is evident as inconsistencies in the literature point to studies that have reported no effect or negative effects of technology use on achievement when compared to not using technology (Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011; Wentworth & Middleton, 2014). Furthermore, current technology effectiveness meta-analyses in statistics education literature have generally used a broad criterion to select primary studies and potential moderator variables (study and methodological characteristics). In accordance with Slavin's (1995) observation of social science meta-analyses, the moderators examined have often been replications of those previously examined in the literature (Sosa et al, 2011), with few examining unique variables related to learner-centered characteristics and pedagogical approaches (Larwin

& Larwin, 2011; Sosa et al, 2011), Meanwhile, experts and scholars call for better quality research on technology effectiveness using approaches that go beyond evaluating technology features (Roblyer, 2005; Schrum et al., 2007). Rather, there is a call for research that provides evidence about the synergies between technology, pedagogy, and content that influence the achievement of learning goals in post-secondary statistics education (Moore, 1997; Pearl et al., 2012). Yet, no meta-analysis studies in statistics education have examined the influences of the interactions of all three.

While meta-analysis has been used to evaluate potential moderators of the effect under study, concerns have also been raised about the validity of conclusion from analyses using findings from primary studies that differ in quality (Hunter & Schmidt, 2004; Slavin, 1995). Given this, a best-evidence meta-analysis approach has been proposed that adds rational to the traditional meta-analysis approach (Clark, 1985; Dochy, 2003; Slavin, 1995). According to proponents of this method, the best-evidence approach goes beyond making conclusions solely based on the analysis of effect sizes (Dochy, 2003; Slavin, 1995). Rather, conclusions are drawn based on the best-evidence from a comprehensive review of quality primary studies that have the most substantive and methodologically sound characteristics. This also includes a discussion of methodological issues identified in the synthesis of studies found in the literature (Dochy, 2003; Slavin, 1995).

Current meta-analyses on technology effectiveness fall short of providing conclusions that bridge the existing gap from research to practice (Roblyer, 2005; Spector et al., 2014; Tishkovskaya & Lancaster, 2012). This may be due to the lack of common methodological and theoretical approaches to the selection of variables relevant to

technology-enhanced instruction (Roblyer, 2005; Schrum et al., 2007; Sosa et al., 2011) and to guide research efforts (Kennewell, 2001; Roblyer, 200). Given this, no known meta-analysis in statistics education assessing technology effectiveness has explicitly implemented a theoretical or conceptual-based framework approach to guide the selection of potential moderators.

According to Roblyer (2005), more quality technology effectiveness studies are needed that address methodological weaknesses of past research and provide direction for future research. Thus, this signals a need for a meta-analysis study on technology effectiveness in statistics education that establishes and employs a framework grounded in theoretical and learning principles to guide the selection of potential moderator variables. Additionally, the investigation should address the complexity of the learning environment by taking into consideration the interactions that occur between technology, pedagogy, and content. Finally, the study should use a best-evidence approach to analyze findings to inform the most effective instructional practices in technology-enhanced environments that support the achievement of learning outcomes in statistics education.

Purpose statement

The current study has three primary aims. First, is to develop a theoretically based conceptual framework to guide the selection of moderator variables. This will occur from a synthesis of literature on the effectiveness of technology use on student achievement compared to not using technology in post-secondary introductory statistics education. Second, is to employ a best-evidence meta-analysis to identify to what extent the synergies between instructional elements related to technology, pedagogy, and content impact students' statistical achievement. Third, is to provide a critical appraisal of the

quality of methodologies employed in the literature and use meta-analytic findings of exemplar studies to recommend the most effective evidence-based strategies for integrating technologies to support students' statistical achievement.

Research questions

1. What is the overall magnitude of the effect of using technology on statistics achievement?
    a. Are there statistically significant variations in the estimated mean effects of using technology on statistics achievement across studies?
2. To what extent do 24 study characteristics associated with phases of instructional design, moderate the effect of using technology on statistics achievement?
3. To what extent are implementation phase elements associated with interrelations between technology, pedagogy, and content predictors of the effect of using technology on statistics achievement?
4. To what extent do report or methodological characteristics of primary studies moderate the effect of technology use on statistics achievement?
5. To what extent is the quality of primary studies a moderator of the effect of using technology on statistics achievement?

Justification

Given the complexity and diverse nature of research in the field, the challenge remains that statistics educators are not well-versed on how to optimize the use of technology to teach learners challenging statistical concepts (Hassad, 2009; Pearl et al., 2012; Tishkovskaya & Lancaster, 2012). According to Sosa et al. (2011), educators are more interested in knowing the best strategies for integrating technologies than merely

knowing that they provide benefits to learning. Additionally, as the costs of technology adoption for teaching and learning can vary (Roblyer, 2005), the current findings could potentially facilitate the decision-making process of policy makers, administrators, and faculty. This can be accomplished by informing the selection of appropriate technologies, their use, and potential benefits to achieve established learning goals (Cobb & McClain, 2001; Chance, Ben-Zvi, et al., 2007; Lajoie, 1997).

Additionally, the current study seeks to add to current research by further explicating past findings using a theoretically-grounded conceptual framework that is based on instructional design, TPACK, and constructivist learning principles. According to Tishkovskaya and Lancaster (2012) "In order to determine whether innovative teaching methods are effective, a link to a theory or theories of learning can be the instructor's most powerful tool in understanding and changing practice" (p.11).

The frameworks used could also provide practical implications for curriculum design and effective instructional planning when integrating technologies in post-secondary introductory statistics education with a focus on the interactions between technology, pedagogy, and content matter that result in enhancing student learning. This is especially important as research priorities suggest the need for identifying the most optimal ways to use technology, given the diverseness of the field in the areas of content, pedagogical practice, and technology use. Therefore, through conclusions made from best-evidence findings, the study could provide relevant information to guide educators' most effective integration of technology to achieve instructional goals.

Furthermore, the current findings could potentially address inconsistencies in conclusions drawn about the impact of technology use in statistics education and inform

future primary research toward measuring relevant variables and testing the applicability of the proposed model in predicting the achievement of learning outcomes when assessing the effectiveness of technology integration in statistics education.

Finally, the methodology employed could potentially direct future meta-analytic research in the field toward enhancing the applicability of research findings by using a common theoretical framework to guide the selection of moderators used to explain differences observed.

Definitions of terms

Educational technology: According to the Association for Educational Communications and Technology (AECT), educational technology is defined as the "disciplined application of scientific principles and theoretical knowledge to enhance human learning and performance" (Spector, 2008, p. 820). Additionally, it refers to "the application of scientific know-how and tools or equipment" (Spector, Merrill, Elen, & Bishop, 2014, p. 6).

Educational technology research: Educational technology research is not only concerned with the important attributes of technologies (what) but also applies theoretical knowledge to understand the "how" and "why" different types of technologies enhance student learning (Spector et al., 2014).

Extent of Risk of Bias: Due to variations in the design, methodologies, and execution of primary studies, an assessment of the extent of risk of bias inherent in studies is necessary to evaluate the validity of studies and the meta-analysis conclusions made. According to the *Cochrane Handbook for Systematic Reviews of Interventions*, an evaluation of risk of bias informs the extent of risk in overestimating or underestimating

(bias) meta-analysis results (The Cochran Collaboration, 2011). Additionally, extent of risk of bias assessment will allow for inferences to be made about the quality of studies included in the meta-analysis. The Cochran Collaboration (2011) suggests that "risk of bias" should be assessed using a tool that assesses the methodological quality of studies. Based on recommendations and examples outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Cochran Collaboration, 2011), the current study uses an assessment tool (risk of bias scale) that assesses methodological features associated with threats to validity (internal, external, implementation fidelity, construct, and statistical validity) across primary studies. The scale uses risk of bias ratings that include "low", "unclear", and "high" risk of bias. These threats of validity correspond with those relevant to research concerned with assessing the effectiveness of the use of technology on student achievement in the classroom. Additionally, adaptations of risk of bias graphical plots are presented from recommendations from the *Cochrane Handbook of Systematic Reviews and Intervention* (Cochran Collaboration, 2011).

Instructional design: Gagne (1974) describes instructional design (ID) as "a body of technical knowledge about the systematic design and conduct of education, based upon scientific research" (p. 3). Though various ID models exist, each encompasses four general components involved in the design of instruction which include: Analysis, Design, Development, Implementation, and Evaluation. Within each component, instructional and learning activities are executed that align with the learning goals and objectives. The current meta-analysis study examines the uses of technologies in primary studies through identifying the instructional design characteristics that have been

implemented and contribute to supporting student learning (Gagne, Wager, Golas, Keller, & Russell, 2005)

Instructional elements: In this study, instructional elements (IE) are the 24 characteristics of the instructional environment related to content, pedagogical practice, and technology use that theoretically form a synergy to facilitate learning (Cobb & McClain, 2004; Schmid et al., 2014S). These are directly associated with the design of instruction and principles of learning in technology-enhanced environments. Thus, instructional design (ID) and Technological, Pedagogical, and Content Knowledge (TPACK) serve as frameworks for identifying the essential elements of the instructional environment in individual primary studies.

Meta-analysis: According to Glass (1976), a meta-analysis is the statistical analysis of results obtained from a large group of individual studies measuring the same phenomenon with the intent of integrating the findings.

Primary Studies: Any empirical research found in the literature and which are candidates for inclusion in the current meta-analysis are referred to as primary studies.

Student statistics achievement: In the current study, statistics achievement relates to students' cognitive knowledge gained in the subject area of introductory-level statistics. Furthermore, it is a learning outcome as a result of integrating technology use to support teaching and learning of statistics. Specifically, statistics achievement encompasses objective measures of established learning outcomes for introductory level post-secondary statistics courses. Across primary studies, these outcomes are reported in a variety of ways including course grades, exam grades, projects/assignment grades, cognitive assessment tests associated course grades, exam grades, and

projects/assignments. Furthermore, categorizations of learning outcomes are associated statistics content covered as identified in the literature (e.g. fundamental statistical concepts) (GAISE College Report ASA Revision Committee, 2016).

Statistics education research: According the Research Advisory Board of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), statistics education research is defined as research designed to inform pedagogical practice for classroom application and to guide future research in the field with new research questions to examine (Zieffler et al., 2008). Furthermore, the goal is to advance teaching strategies that enhance learning outcomes (Zieffler et al., 2008).

Statistical literacy: Statistical literacy is concerned with the basic ability to read, understand, interpret, predict, and critically think about statistical information and argue claims that are made (Ooms & Garfield, 2008; Sharma, 2017).

Statistical Reasoning: Statistical reasoning is how individuals reason and make sense of provided statistical information (Garfield, Chance, Poly-San, & Obispo, 1999).
Statistical Thinking: Statistical thinking is associated with learners' ability to think quantitatively and can be regarded as stemming directly from statistical reasoning (R A Hassad, 2009).
Technology: Technology in this study, is a tool - computer hardware, software, or digital artifact that is either produced commercially or designed in-house and is used to carry-out instructional practices that support student learning of statistics. This includes tools such as graphing calculators, data analysis, graphic, and visualization software, drill and practice tutorials, multi-media, simulation, Internet, communication technologies, and computer-based learning management systems (Lajoie, 1997). Furthermore, technology

use encompasses not only the tool, but also the elements in the instructional context related to pedagogical strategies and content that interact with technology use to support student learning.

Technology affordance and constraints: Technology affordances relate to the attributes of technology-enhanced settings that promote action; whereas, constraints relate to conditions and relationships shared by attributes that control the conditions in which actions can take place (Kennewell, 2001). In environments where technology is used, an individuals' ability to use technology to accomplish objectives is dependent on their knowledge, skills, and understanding (Kennewell, 2001).

Delimitations

The delimitations of the current study consist of the researcher's choice of selection criteria that excludes studies from the meta-analysis. This refers to the exclusion of studies conducted prior to 1998 that meet the inclusion criteria. Additionally, among those published within the inclusionary period of 1998 – 2018, excluded are those that do not use an experimental or quasi-experimental research design and those that use a sample of students enrolled in an advanced post-secondary statistics course. These decisions are justified based on the goal of the study, which is to assess the effectiveness of instructional strategies that can inform best-practices for technology integration in post-secondary introductory statistics education. According to Cooper & Hedges (1994), one approach to enhance the construct validity of a meta-analysis is to place restrictions on the selection criteria to the use of studies that use experimental or quasi-experimental designs. Furthermore, this will enhance the assessment of the true effect of technology use based on the most rigorous research designs, also enhancing the statistical conclusion

validity of findings reported. The disadvantage of this approach, however, is that it might potentially limit the size of the population from which samples are drawn. Furthermore, as students' ability to grasp fundamental statistics concepts is essential at the post-secondary introductory-level, limiting the context to technology use at the lower level of statistics education is deemed appropriate and significant for identifying practical considerations for using technology to support learning early on.

Assumptions

The application of the ADDIE (Analyze, Develop, Design, Implementation, Evaluation) model in the current study is not for the purpose of confirming or testing a definite sequence of steps for designing or implementing elements of instruction when using technology for supporting statistical learning. That can only be accomplished through the deliberate planning and design of instructional activities, implementation, and direct measurement of related constructs to evaluate the effectiveness of the ID model in the unique instructional context. In the current study, the coded elements represent *synthesis generated* evidence as described by Cooper (2017). These are evidence that have not been directly measured, thus, causality cannot be conferred. However, *synthesis generated* evidence are useful for capturing variations in procedures across primary studies and to test relations not previously examined (Cooper, 2017).

Furthermore, the ADDIE and TPACK facets used in the study serve as frameworks and references to guide the inclusion of instructional design elements related to content, pedagogy, and technology in technology-enhanced statistics education learning environments – as substantiated in the literature. They are useful for outlining and prescribing the constructivist instructional activities that lead to the achievement of

learning outcomes in introductory statistics education – as substantiated in the literature. The developed conceptual framework then provides a guide from which meta-analysis can be conducted (using a theoretically/conceptually grounded approach) for identifying relevant variables (moderators) that can provide an explanation for the observed differences in effect sizes across primary studies.

CHAPTER II – REVIEW OF THE LITERATURE

Addressing the gap between theory and practice in post-secondary statistics education involves identifying the optimal instructional practices for using technology to support teaching and learning. Accomplishing this requires overviews of the development of technology-based reform in statistics education and the literature on technology effectiveness that highlights the roles of technology, pedagogy, and content in supporting cognitive achievement. Furthermore, findings from existing primary and meta-analysis studies provide insight through empirical investigations on the effectiveness of technology use in education and statistics education. Meanwhile, an examination of the current state of meta-analysis research reveals points for consideration for improving future technology effectiveness research. Finally, the application of theoretical frameworks consisting of Instructional Design (ID), Pedagogical, Technological, and Content Knowledge (PTCK), and constructivism provide lenses for evaluating primary empirical research, as well as for employing a best-evidence meta-analysis approach to inform best-practices in using technology to support statistical achievement in post-secondary introductory statistics classrooms.

Development of technology-based reform in statistics education

The prominence of statistics as a practice has early beginnings in the mid-eighteenth century, when it was primarily used by arithmetic politicians who collected and analyzed data to make sense of and to make predictions about observations in the society, the population, and the economy (Ben-Zvi et al., 2017; Hassad, 2009). Later, the field experienced growth internationally as a scientific discipline and contributions from leading organization such as the Royal Statistical Society (RSS), the American Statistics

Association (ASA) and the International Statistical Institute (ISI) led to the shaping of the field with a commitment and command for training and research in statistics education (Hassad, 2009). By the early 20th century, statistics became primarily a vocational area of study geared towards practicing scientists with an emphasis on mathematical computations (Aliaga et al., 2012). By mid-century, it was established as an academic discipline for aspiring scientists – with a focus in the content area of probability (Aliaga et al., 2012). Teaching practices focused on developing students' knowledge, methodological skills, and computational abilities (Tishkovskaya & Lancaster, 2012).

According to Aliaga et al. (2012), the publications of *Statistics* by David Freedman, Robert Pisani, and Roger Purves and *Statistics: Concepts and Controversies by David S. Moore*, both in 1978, led to the introduction of statistics as an introductory course in academia. Statistics became an academic discipline taught in all levels of education (from primary to post-secondary), as well as a required introductory course for many students in a variety of disciplinary fields at the post-secondary level (Cobb, 2007; Hogg, 1991; Everson et al., 2008; Tishkovskaya & Lancaster, 2012). Later, statistics education experienced a notable shift in its content and pedagogy (Cobb, 1992; Tishkovskaya & Lancaster, 2012). This shift was motivated by a movement of reform largely focused on improving learning outcomes in introductory level statistics courses (Aliaga et al., 2012; Cobb, 1992; GAISE College Report ASA Revision Committee, 2016; Pearl et al., 2012).

Challenges in learning statistics. Following a meeting of statisticians from leading organizations in the field, Hogg (1991) pointed toto challenges associated with pedagogical practices and students' lack of preparation. Hogg insisted that mathematical

concepts should not be the foundations from which introductory statistics courses are taught. Yet, mathematical and computational approaches continued to be the bases used for teaching statistical concepts thereby, contributing to challenges in both teaching and learning (Moore, 1997; Schuyten & Thas, 2007). Additionally, empirical findings in teaching and learning statistics revealed various challenges associated with learners' cognitive inabilities to grasp key fundamental concepts at the introductory level (Everson, Zieffler, & Garfield, 2008).

The concerns raised by Hogg (1991) contributed to the urgency in improving statistics education, with statistical thinking and statistical reasoning as major concepts that should be taught (Garfield & Ben-zvi, 2008). At the same time, however, given the diversity of learners taking introductory-level statistics, it was found that many lacked the pre-requisite knowledge or had no prior exposure to the content, often associating the subject to a mathematics course and thus leading to students' negative attitudes and raised anxiety (Everson, et. al, 2008; Hassad, 2009). Castro Sotos, Vanhoof, Van den Noortgate, and Onghena (2007) conducted an exploration of empirical literature published from 1990 to 2006, and found several misconceptions held by students related to fundamental concepts of sampling distributions and variability, central limit theorem, hypothesis testing, significance levels, statistical significance, p-value interpretation, and confidence intervals. Although students grasped concepts enough to pass an introductory-level statistics course, they struggled with the same concepts when faced with them in successive courses, demonstrating that they still lacked a clear understanding of fundamental concepts (Cobb, 1992; Garfield & Ben-Zvi, 2007). Chiesi and Primi (2010) suggested that these difficulties are even more pronounced among students with

qualitative academic backgrounds taking introductory statistics courses. Furthermore, the resistance and anxiety experienced by undergraduate students toward learning statistics makes teaching statistics challenging, leading to student underperformance in statistics (Chiesi & Primi, 2010; Lloyd & Robertson, 2012).

Challenges in teaching statistics. In addition to the challenges faced by learners, the interdisciplinary nature in which statistics is taught makes it unique, having different approaches in how instruction is carried out, differing areas and degrees of focus related to content covered, varying access to instructional resources, a diversity of learners with different levels of cognitive ability and motivation concerning the subject, and instructional contexts that vary in size of groups of learners taught (Cobb, 1992; Garfield & Ben-Zvi, 2007; Zieffler et al., 2011). Given the diverseness of the learning environment, statistics educators are faced with difficult instructional tasks of presenting appropriate content to meet the learning needs of students while ensuring that expected learning goals and outcomes are being achieved (Garfield, 1995). The realization of the diverse contexts in which the subject is taught required changes using new pedagogical approaches and the integration of innovative educational technologies to support instructional delivery (Tishkovskaya & Lancaster, 2012).

Furthermore, with increasing innovations in educational technologies, statistics educators are faced with a variety challenges associated with selecting the most appropriate technologies, costs associated with technology adoption, learning how to use these tools and deciding on the most effective method of integration in order to yield achievement of student cognitive and affective learning outcomes (Cobb & McClain, 2001; Chance, Ben-Zvi, et al., 2007). Additionally, institutional policies, facility

constraints, educators' lack of knowledge, skills, and ability in using technology are some factors that may inhibit technology integration in the classroom (Buabeng-Andoh, 2012; G. Cobb, 1992; Kim, Kyu, Lee, Spector, & Demeester, 2013; Pearl et al., 2012; Tsai & Chai, 2012). Similarly, the use of technology for teaching and learning requires that learners possess the skills and abilities to utilize the tools in order to achieve intended outcomes (Bates, 2015; GAISE College Report ASA Revision Committee, 2016; Lowerison et al., 2006; Nguyen et al., 2016). The extent to which this occurs involves a joint effort of students' self-directedness, as well as guidance and scaffolding provided by the instructor (Carver et al., 2016; Garrett, 2016; Schuyten, G., & Thas, O., 2007; Tishkovskaya & Lancaster, 2012; Garrett, 2016; Peterson, 2016; Lloyd & Robertson, 2012).

Reform in teaching and learning statistics. The early call for improvements in statistics education highlighted by Cobbs (1992) and his colleagues laid the groundwork and paved the direction for reform efforts toward improving outcomes in statistics education through changes in content, pedagogy, and technology integration. During the 1960s and 1970s, the evolution of computers changed the way statistics instruction could be delivered. In his seminal work, *Teaching Statistics*, Cobb (1992) acknowledged a shift in three areas of statistics education within the previous two decades related to "technique, practice, and teaching" (p. 4). Moore (1997) described the reform in terms of changes in content (more data analysis, less probability), pedagogy (fewer lectures, more active learning), and technology (for data analysis and simulations). This led to the use of innovative tools such as drill and practice tutorial, multimedia, simulation, and visualization software (Aliaga et al., 2012; Larwin & Larwin, 2011).

Furthermore, a focus on mathematical hand-calculations gave way to graphical methods and display of data (Schuyten & Thas, 2007). Technological advancements afforded new graphical methods to display data, enhanced capabilities of analysis software, and the availability of tools to facilitate data exploration and manipulation lead to new ways of teaching theoretical concepts (Cobb, 1992). These enhanced capabilities of analysis software facilitated new ways of teaching theoretical concepts (e.g. hypothesis testing). The prevalent use of technology in statistics education was further evident in the findings of a 2001 survey conducted by Bratz and Sabikuj that reported an increase in technology adoption by universities in introductory-level statistics courses from 50% in 1982 to 80% after more than two decades (Larwin & Larwin, 2011).

To address concerns with students' achievement of learning outcomes, Garfield (1995) insisted on a focus on theories of learning to guide changes in instructional practices in introductory level statistics education. Constructivist approaches to instruction were urged by those concerned with the achievement of these course outcomes (Moore, 1997; Garfield & Ben-zvi, 2007). Furthermore, Moore (1997) acknowledged the social context in which statistics education had evolved and insisted that this realization should be accompanied with changes in what is taught and how it is taught – with technology serving an influential role.

Current state of statistics education. Recent developments in improving statistics education have been influenced by the contributions of professionals, researchers, and leading organizations in the field (College Report ASA Revision Committee, 2016; Pearl et al., 2012). In an effort to guide statistics educators in addressing the challenges faced in teaching and learning statistics, organizations such as the National Science Foundation

(NSF) and the American Statistics Association (ASA) called for reform through suggested teaching practices focused on enhancing statistical thinking, statistical literacy, application, use of data, and use of technologies that provide opportunities for increased engagement and active learning to occur (GAISE College Report ASA Revision Committee, 2016). ASA published the Guidelines for Assessment and Instruction in Statistics Education College (GAISE) 2010 and 2016 reports to address reform in introductory level statistics and upper college level courses. The 2016 report highlights six main recommendations for improving and enhancing statistics education which include:

> 1) a focus on statistical thinking, 2) a focus on students' conceptual understanding, 3) the use of real data within context, 4) activities that support active learning, 5) the use of technology for data analysis and exploration of concepts, and 6) the use of assessments to evaluate student learning and provide feedback for improvement (GAISE College Report ASA Revision Committee, 2016, p. 3).

Moreover, the enhancement of students' cognitive learning outcomes has focused on fundamental concepts (the "Big Ideas") of probability, variation, randomness and statistical competencies related to statistical literacy, thinking, and reasoning (Cobb & McClain, 2001; GAISE College Report ASA Revision Committee, 2016; Garfield & Ben-Zvi, 2004a; Pearl et al., 2012). More emphasis has been placed on these learning outcomes over mathematical operations and procedures (Garfield and Ben-Zvi, 2008).

The inclusion of technology in the recommendations set by GAISE have supported a shift from traditional instructional practices (e.g. lecture-based) to non-traditional learner-

centered methods that allow students to be active participants in learning (Aliaga et al., 2012; GAISE College Report ASA Revision Committee, 2016). According to Roseth et al. (2008) traditional methods of teaching statistics are not as effective as those that support engagement and provide collaborative opportunities to learn. This has given way to the adoption of various instructional delivery strategies in introductory-level college statistics courses. With the abundance of free online statistics resources, educators and learners have a variety of tools at their disposal that are useful in supporting teaching and learning activities (Chance, Ben-Zvi, et al., 2007). Furthermore, assessment practices have been emphasized due to the importance of monitoring the effectiveness of teaching and learning practices in order to improve the achievement of learning outcomes (GAISE College Report ASA Revision Committee, 2016).

Technology effectiveness literature

Research interests have been fueled by increased innovations in educational technologies, the growing acceptance and use of technologies by post-secondary educators and learners, and accountability pressures associated with the high costs of technology implementation that requires that decision makers understand its benefits to teaching and learning (Chance, Ben-Zvi, et al., 2007; Lowerison et al., 2006; Roblyer, 2005; Schrum et al., 2007; Thompson, Bell, Schrum, & Bull, 2008). Technologies, also referred to as information and communication technologies (ICTs), computer assisted instruction (CAI), computer mediated communication (CMC), or computer based instruction (CBI) are increasingly being used in all levels of education (Hsu, 2003; Kennewell, 2001; Kulik, Kulik, & Shwalb, 1986; Roseth, Akcaoglu, & Zellner, 2013). The prominence of technology use in post-secondary education has been met with

research assessing its effectiveness. One of the aims of technology effectiveness studies is to assess if using technology enhances student achievement (or cognitive or affective learning outcomes) when used to support teaching and learning in the classroom (Borokhovski, Bernard, Tamim, Schmid, & Sokolovskaya, 2016; Chance, Ben-Zvi, et al., 2007; Lajoie, 1997; Roblyer, 2005; Schmidt et al., 2014; Thompson et al., 2008). These studies have often been categorized under educational technology research, while spanning across disciplinary areas (Morrison & Ross, Steven, 2014; Warren, Lee, & Najm, 2014), including statistics education (Garfield & Ben-zvi, 2007). In general, research has focused on comparing different types of technologies used in the classroom on learning outcomes, as well as experimental (using technology) versus control conditions (not using technology) (Schrum et al., 2007). Fewer primary studies have used randomized control research designs compared to quasi-experimental and qualitative designs (Ross & Morrison, 2014). Meanwhile, an increasing presence of meta-analysis studies have examined the overall impact and moderating factors that influence the effectiveness of using technologies to support teaching and learning (Zieffler, 2018). Furthermore, cognitive outcomes (e.g. student achievement) have most often been the measurement used for substantiating the impact of technology use on learning (Ross & Morrison, 2014). In large, studies have reported positive findings on the effectiveness of using technology as a medium for transforming and affecting learning (Archer et al., 2014; Lakhana, 2014; Robinson et al., 2009; Tamim et al. (2011).

Technology's role in supporting learning. The main goal of educational technology research is to understand how the medium (technology) and the method (instructional strategies) interact to enhance learning (Bernard et al., 2009; Morrison &

Ross, Steven, 2014). However, historically, there have been diverging views within the literature about the role that technology plays in influencing learning (Roblyer, 2005; Schmid et al., 2014). On one hand, advocates claim the use of technology is effective in transforming and enhancing learning (Archer et al., 2014; Kozma, 1994; Lakhana, 2014; Robinson et al., 2009; Tamim et al., 2011). Meanwhile, opposing views have argued that technology is nothing more than a vehicle (medium) for transporting knowledge and alone, does not influence gains in learning (Clark, 1985; Clark, 1994). Also criticized have been claims that technologies provide greater advantages in student achievement over traditional instructional methods (Clark, 1985; Clark, 1994). Yet, there is consensus in that pedagogical strategies employed and content covered contribute to transformational learning in technology-enhanced learning environments (Bernard et al., 2009; Clark, 1983; Moore, 1997; Schmid et al., 2014S; Schrum et al., 2007; Tamim et al., 2011). Kozma (1994) added that the relationship between the use of media and learning can be explained by examining the interactions that occur between cognitive processes and characteristics of the learning environment. Commonly agreed is that it is the interactions among technology, pedagogy, and content that contribute to transformational learning in technology-enhanced learning environments (Bernard et al., 2009; Cobb & McClain, 2004; Schmid et al., 2014S; Schrum et al., 2007; Shulman, 1986; Tamim et al., 2011).

Effectiveness of technology use in statistics education. Synthesizing the studies that assess the effectiveness of technology use in statistics education reveals a collection of empirical studies that are diverse in research scope, methodologies employed, and outcomes measured. Zieffler et al. (2008) noted that the landscape of empirical research

on teaching and learning in statistics education comprises a variety of methodological approaches, participants sampled (primary-level to post-secondary-level students and professionals), research questions addressed, and outcomes measured. This diversity has made it difficult to establish clear conclusions about the overall effectiveness of reform-oriented pedagogical strategies (Hassad, 2009).

Furthermore, Cobb (2007) remarked that though research in the area of teaching and learning statistics has grown, the interdisciplinary nature of the field has led to diversity in research as each discipline has adopted unique research methods, perspectives, and inquiry focus. For example, statistical reasoning, statistical thinking, and statistical literacy are outcomes heavily researched by those in the field of psychology and mathematics. In contrast, statistics educators have been mainly concerned with the effective use of technology in achieving specific learning goals, enhancement of students' attitudes towards statistics, and reduction of statistics anxiety (Ciftci, Karadag, & Akdal, 2014; Garfield & Ben-zvi, 2007).

Many of these studies have been conducted in a classroom setting, emphasizing the instructional strategies implemented, focused on the use of a particular tool or multiple tools, and measuring a variety of outcomes (Garfield & Ben-zvi, 2007; Tishkovskaya & Lancaster, 2012; Zieffler et al., 2008). Some of the outcomes examined have included multiple choice exam (Basturk, 2005; Mclaughlin & Kang, 2017), statistical problem set (Lloyd & Robertson, 2012), final exam (Phillips & Phillips, 2016), The Comprehensive Assessment of Outcomes in a First Course in Statistics (CAOS) and topic scales from ARTIST (Mcgowan & Gunderson, 2010), The Statistics Achievement Scale (Ciftci, Karadag, Akdal, & Pinar, 2014), as well as course grades, assignments, etc.

A large number of these studies have been based on the researcher's evaluation of their own class at a single point in time or across multiple classes (Garfield & Ben-zvi, 2007).

Furthermore, evidence in the literature points to the idea that the thoughtful design and sequencing of activities and the use of technology can improve statistics students' reasoning and understanding of fundamental concepts in statistics (e.g. distribution, variation, etc.) (Garfield & Ben-zvi, 2007; Zieffler, 2008). Meanwhile, several authors note that identifying effective tools for learning, as well as those for guiding and monitoring students' use of technology are essential for the appropriate assessment of learning (Garrett, 2016; Lloyd & Robertson, 2012; Peterson, 2016; Schuyten & Thas, 2007; Tishkovskaya & Lancaster, 2012).). Moore (1997) emphasized the need for reform focused on the content covered in introductory statistics education, advocating that technology creates a synergy with content and pedagogy that leads to effective instruction. Similarly, Scheaffer (1997) insisted that the use of technology to support teaching and learning of content should include students' use of technology to explore concepts of statistical inference.

Types of technologies used. Iiyoshi, Hannafin, and Wang (2005) argued that despite the push for constructivist student-centered technology-enhanced learning environments, certain tools can present a cognitive burden on student learning. However, when used appropriately, these tools are useful in scaffolding the learning process (Iiyoshi et al., 2005; Schmid et al., 2014), as well as enhancing cognitive processes and supporting the creation of students' knowledge (Iiyoshi et al., 2005). The authors recognized that technological tools support both cognitive functions and the achievement of learning goals. Furthermore, according to Chance, Ben-Zvi, Garfield, and Medina

(2007), while technology-enhanced instruction should be focused on the content matter, the selection of the appropriate tools should be guided by the learning goal. Addressing the role of cognitive tools in supporting student learning, Iiyoshi, Hannafin, and Wang (2005) classified them according to five categories: information seeking, information presentation, knowledge organization, knowledge integration, and knowledge generation. In addition, the authors described the different goals associated with cognitive tools which include: automation of calculations, emphasis on data exploration, visualization of abstract concepts, simulations as a pedagogical tool, investigation of real-life problems, and provision of tools for collaboration and student involvement.

The usefulness of technology in statistics education is described as to either facilitate/enhance problem solving or to alter the conceptualization or understanding of how an individual approach solves a problem (Cobb, 2007, Pearl et al., 2012). The technologies used come in a variety of formats ranging from commercial-based (propriety) to teacher-produced online learning tools. Tools include graphing calculators, television, computational software, videos, statistical software, multimedia tools (Moore, 1997; Garfield & Ben-Zvi, 2007), vodcasts and podcasts (Lloyd & Robertson, 2012), learning management systems (LMS) and Wiki's that support student-teacher communication, collaboration with peers, feedback and reflection (Chance, Ben-zvi, et al., 2007), tutorials (e.g. drill and practice, screencasts) (Chance, Ben-zvi, et al., 2007; Lajoie, 1997). These technologies have been categorized as tools to: deliver instruction (e.g. non-traditional, fully-online, hybrid and flipped course formats), support instruction (e.g. simulation, real data, screencast tutorials), and support learning (e.g. visualization,

applets, web resources) (College Report ASA Revision Committee, 2016; Poly & Obispo, 2007; Robinson et al., 2009).

Garfield and Ben-Zvi (2004) further distinguished among types of technologies used in statistics education. These include commercial statistical packages used for data analysis and displaying visual representations of data (e.g. spreadsheets); data analysis software that provide capabilities for both simulations and visual representation of data that can be manipulated; educational data analysis tools that support data analysis inquiry and graph plotting (e.g. Fathom); web or computer-based applets; stand-alone simulation software (e.g. SIM); the Internet; CMC technologies that are used to support learning in face-to-face and distant education environments (e.g. online forums, online communities, email); as well as technologies useful for developing students' statistical reasoning (e.g. online data sources for data exploration) (Garfield & Ben-Zvi, 2004).

Primary and meta-analysis empirical investigations

Primary studies

Both quantitative and qualitative studies have provided evidence and have described the ways that technology can be used to deepen students' understanding of statistical concepts, and to address misconceptions held by learners (Chance, Ben-Zvi, et al., 2007; Roblyer, 2005; Mcgrath, 2014) . These include studies that examine the development and use of a tool or several tools in the classroom, compare different tools, describe the instructional activities associated with technology-enhanced instruction, or assess the effectiveness of a tool or curricular approach on statistical reasoning and other statistics learning outcomes (Borokhovski et al., 2016; Chance, Ben-zvi, et al., 2007; Lachem, 2014; Lloyd & Robertson, 2012; Schwier & Seaton, 2013). Furthermore, while

conducted in classroom settings, these studies have examined technology use in statistics courses taught in a variety of disciplines (ex: statistics, biostatistics, social sciences, business, psychology, etc.) for undergraduate and graduate course, and at introductory and advanced levels.

For example, concerning the development of pre-service teachers' (PST) pedagogical content knowledge of elementary statistics concepts, Francis, Hudson, and Vesperman (2014) examined the influence of integrating technology use (e.g. Tinkerplots across three different problem-based learning approaches - project based learning (PbL), problem solving (PS) and model eliciting activities (MEA). While there were no differences on PSTs understanding across types of learning approach, all groups showed an increase in their understanding of concepts from pre-test to post-test. The authors concluded that the increase in students' understanding supported the use of appropriate technologies and -solving curricular activities with contextualized content in enhancing PSTs statistical literacy. Furthermore, the authors stressed that to enhance students' statistical literacy, technology-rich environments should incorporate opportunities for data exploration that are within the context of authentic problems, structure activities that allow students to identify the tools' computational and analytical functions, and scaffold learning to support the development of students' statistical reasoning.

Similarly, Garfield, DelMas, and Zieffler (2012) evaluated the use of Tinkerplots for modeling, simulation, and inference to develop tertiary-level students' statistical thinking about randomization and resampling. Using two researcher-developed assessment instruments (the Goals and Outcomes Associated with Learning Statistics (GOALS) and the Models of Statistical Thinking (MOST)) to measure introductory-level

course learning outcomes associated with students' statistical thinking and reasoning, results revealed positive learning gains associated with students' statistical inferences using Tinkerplots for modeling and simulation. The authors concluded that findings supported the use of software designed with an understanding of how introductory-level students learn, as well as a curriculum design that allows flexibility in content and pedagogy based on students' learning progression.

Modes of instructional delivery. The presence of non-traditional delivery formats are increasing in statistics education as learning is no longer confined to space and time (Peterson, 2016; Yamagata-Lynch et al., 2015). This has led to the use and availability of online learning tools and resources that can be accessed and used by students at any time to supplement and reinforce classroom teaching (College Report ASA Revision Committee, 2016; Peterson, 2016; Phillips, & Phillips, 2016). According to Boyer et al. (2013), the flexibility offered by online learning supports students' need and development of self-directed behaviors. The technology-enhanced instructional delivery strategies investigated have included (among others) flipped, online, and hybrid classrooms, along with pedagogical strategies related to cooperative, collaborative, and project-based learning (Chance, Ben-Zvi, et al., 2007; GAISE College Report ASA Revision Committee, 2016; Larwin & Larwin, 2011; Zieffler et al., 2008). Blended learning classrooms (at least 50% online with face-to-face teaching (Schmid et al., 2014)) provide an environment by which traditional classroom instruction time is complimented with the use of multi-media resources to support students' self-directed learning outside of the classroom. This instructional delivery type supplements learners' needs for interaction with teachers, peers, and content as learning occur at any time in the online environment.

Research investigations in educational literature on the effectiveness of blended learning environments compared to a traditional learning environments have found positive teaching and learning outcomes favoring blended learning environments (Gebre, Saroyan, & Bracewell, 2014, Schmid et al., 2014). However, some primary studies in statistics education have reported no differences when comparing blended learning to traditional classroom environments on student performance (Utts, Sommer, Acredolo, Maher, & Matthews, 2003; Ward, 2004). Furthermore, it has been noted that designing instruction should assess the appropriateness of content taught relative to the conditions or context in which learning occurs (Cobb & McClain, 2001; Khalil & Elkhider, 2016; Wessa, Rycker, & Holliday, 2011). As an extension of blended learning and with the onset of technology advancements, the flipped classroom format has become of greater interest as it allows teachers to reverse the traditional teaching format. Content traditionally covered in lectures are adapted to video or through online media such that students explore, engage, and are presented with course content and materials outside of class and with class time devoted to the practical application of content presented (McGraw & Chandler, 2015; Robinson et al., 2009). A relatively small number of studies (mostly conducted in higher education) have evaluated the effectiveness of this type of learning format (Mclaughlin & Kang, 2017; Peterson, 2016) and have reported findings of positive outcomes or no difference compared to traditional formats (Garfield & Ben-zvi, 2007). As flipped classrooms can be designed differently, studies have employed varying approaches, ranging from the use of outside of class textbook reading, to the use of technologies to support the presentation of content. Additionally, other resources used

to supplement flipped formats have included online learning tutorial resources (e.g. Khan Academy) or journal articles (McGraw & Chandler, 2015; Mclaughlin & Kang, 2017).

In a quantitative study, Wilson (2013) evaluated the effectiveness of flipping a lecture statistics course using textbook reading and reading quizzes to present content outside of the classroom and activities working with problem sets for in-class learning. Although overall, students in the flipped classroom performed better than their lecture course counterparts, a large number of students were not satisfied with the textbook mode of presentation used. Furthermore, a quantitative study conducted by Strayer (2012) using a flipped format in a statistics course assessed the use of a tutoring system outside of class, supplemented with in-class activities. Findings revealed that students were dissatisfied as the content covered in-class did not align with the material covered out-of-class. These findings supported the importance of aligning course content with the appropriate use of technologies to support teaching and learning (Chance, Ben-zvi, Garfield, & Medina, 2007; Moore, 1997; Wessa, Rycker, & Holliday, 2011).

Addressing this concern, Peterson (2016) evaluated the effectiveness of teaching and learning statistics in a flipped course learning environment that incorporated online outside of class learning activities compared to a traditional classroom (lecture) environment on student performance and students' perception of their learning experience. Using a sample of 43 university students in an introductory statistics course, findings revealed that on average, students in the flipped class had higher achievement on the final exam than those in the traditional classroom, as well as reported greater satisfaction with the course overall. Similarly, Mclaughlin and Kang (2017) examined the affect of a shortened (two-weeks – nine meetings) interactive foundation biostatistics

flipped classroom model course for health science education doctoral students on student achievement and course satisfaction. The course was designed using constructivist principles and students completed online pre-course modules (instructor-developed) to supplement in-class activities. Using a single cohort pre-test, post-test design, analysis revealed that students performed higher on the final examination than at pre-test, indicating the usefulenss of short-course formats for enhancing students' understanding of fundamental biostatistics concepts. Furthermore, course evaluations completed by students revealed that students' motivation increased as a function of their satisfaction with the level of engagement, learning content, and usefulness of the course. Mclaughlin and Kang (2017) noted that future research should examine the relationship between instructional design aspects of the flipped format and effective pedagogical practices that lead to learning gains.

Technologies for supporting instruction. Technologies developed by teachers such as vodcasts, podcasts, and screencast tutorials have been used in statistics education to promote interactive learning, and provide an enhanced learning experience (Lloyd & Robertson, 2012).This occurs as statistical concepts are demonstrated through a multimodal platform that presents information in both audio, video and text formats, supports reduced cognitive loading and enhanced cognitive processing and allows for deep learning can occur with clear communication of information as individuals engage in self-paced learning Mayer (2014). In their study, Lloyd and Robertson (2012) investigated the of using a screencast video tutorial in an undergraduate statistics course for psychology students. The screencast tutorial was used as a supplement to classroom teaching to enhance students' knowledge, application, and interpretation of statistical

concepts. Students were randomly assigned to two conditions – a control text tutorial group and an experimental screencast video tutorial group. Controlling for confounds of math experience, math and computer anxiety, and course grades, the findings revealed that the screencast video tutorial was more efficient and effective in enhancing students' learning of the statistical concepts presented than did the traditional teaching approaches, especially with more complex concepts.

Additionally, clickers (or Action Response Systems) are increasingly being used in general education and statistics education whereby students use wireless hand-held devices to respond to and to ask questions, also allowing for immediate feedback and increased student engagement (Ramesh, 2011). However, amid continuous debate about the effectiveness of clickers (or Action Response Systems) in supporting student learning, Mcgowan and Gunderson (2010) conducted a randomized experiment to investigate how identified features (number of questions asked, placement of questions asked, grades) associated with clicker use affect undergraduate statistics' students' engagement and learning in statistics. The Comprehensive Assessment of Outcomes in a First Course in Statistics (CAOS) and topic scales from the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project website were used to measure student learning. Though no evidence was found for increasing engagement, the authors found that the use of clickers enhanced student learning. was attributed to the careful placement of questions in instructional material and fewer clicker questions presented to students.

Technologies for supporting learning. González, Jover, Cobo, and Muñoz (2010) commented on the large variety of online learning resources that contain topics related to statistics that are available in a variety of multimedia formats from basic simulations to

web-based textbooks. The use of multimedia technologies provides a space for exploration of information, offering visual and audio presentation of content simultaneously and opportunities for learners to engage with simulation technology (Schuyten and Thas, 2007). These types of technologies, used in statistics education, come in a variety of formats ranging from screencast tutorials, simulations, web-based resources and other teacher-produced online learning technologies. According to Mayer's (2014) cognitive theory of multimedia learning, in-depth learning occurs through the simultaneous processing of auditory and visual stimuli, which supports the way the brains functions and leads to active processing of information. Thus, multimedia tools are often positively perceived and are often used by students to engage in learning activities (Garrett, 2016; González et al., 2010; Phillips & Phillips, 2016; Schuyten, & Thas, 2007). These tools support constructivist learning and benefit the teaching and learning of statistics, allowing students to control their learning, engage with information, tackle real-world problems, construct and make-meaning of their own knowledge (Lloyd & Robertson, 2012; Poly & Obispo, 2007; Schuyten & Thas, 2007). Furthermore, the use of simulation programs has been found to support and enhance students' development of statistical reasoning (delMas, Garfield, and Chance; 1999; Lunsford, Rowell and Goodson-Espy, 2006), as well as to have greater effects on learning statistical concepts compared to textbook instructional methods (Lane & Tang, 2000). In contrast, in regard to the effectiveness of tutorials, Aberson, Aberson, Berger, Healy, Kyle and Romero (2000) reported no significant differences on improvements in statistical learning between groups that used an online tutorial and those that used traditional lecture.

Furthermore, Gonzalez et al. (2010) conducted a study using a randomized experimental approach to evaluate the effect of an instructor produced web-based learning tool on improving 121 dentistry undergraduate students' performance in statistics. Students were randomly assigned to a control (traditional problem-solving approach - paper) or a treatment condition (web-based problem-solving approach – e-status). The results revealed that the use of the web-based learning tool (e-status) positively influenced students' learning of statistical numerical operations. Students in the e-status group showed greater improvements in statistical abilities, compared to students in the paper-based group. With regard to simulation tools, although they offer students opportunities to understand statistical concepts (e.g. random processes), Garfield and Ben-Zvi (2007) noted that they are only effective when their use has been carefully planned into instruction.

Multi-media environments allow learners to interact directly with content, supporting cognitive processing of data that help develop learners' statistical thinking abilities that are necessary in solving statistical problems (Schuyten & Thas, 2007). Through the use of multi-media tools, students learning statistics benefit from rich audio/video tutorials (e.g. screencast tutorials, applets) that provide opportunities for learners to scaffold their learning, gain immediate feedback, conceptualize knowledge gained, and enhance students' statistical thinking (Buzzetto-More, 2014; Poly & Obispo, 2007; Schuyten & Thas, 2007). Though students have provided positive feedback about their experience using multimedia tools in learning and being taught statistics, according to (Schuyten & Thas, 2007), conflicting views exist about the need for structure when using these tools, especially during self-regulated learning in a computer-based

environment. Furthermore, technologies have been used to support student-teacher communication. These are referred to as computer mediated communication (CMC) technologies and they support collaboration and student engagement in face-to-face and distance education environments. Examples of tools used include course management systems, online forums, email, etc. (Garfield et al., 2008).

Meta-analysis studies

With beginnings over several decades ago in education and the social sciences, meta-analysis has grown in its acceptance and relevance in various other fields of study (e.g. medical, sciences, psychology, etc.) as an empirical method for assessing the overall impact of interventions and for informing practical decisions and policy making (Borenstein, Hedges, Higgins, & Rothstein, 2017; Cooper & Hedges, 1994; Rosenthal & DiMatteo, 2001; Slavin, 1995). It is a technique whereby statistical findings from studies addressing the same hypotheses about a phenomenon are analyzed (Cooper & Hedges, 1994; Denson & Seltzer, 2010). Also referred to as research synthesis or a quantitative systematic review, it gained prominence in social science research during the 1960s-1970s (Card, 2012: Slavin, 1995).

The most influential use of the technique has been attributed to the seminal works of Gene Glass and his colleagues, at which time the term "meta-analysis" was coined (Card, 2012; Cooper, 2017; Slavin, 1995). In education and social science research, meta-analysis has allowed for the assessment of the overall impact of a variety of phenomena, including implemented educational strategies or programs (Card, 2012; Cooper & Hedges, 1994; Field & Gillet, 2010; Rosenthal & DiMatteo, 2001). This comes with its quantitative distinction for providing empirical evidence: concerning the overall

magnitude of the effect of a phenomenon that explains differences in effect sizes across studies through moderator analyses; and that is generalizable (Cooper & Hedges, 1994; Field & Gillet, 2010). Furthermore, the meta-analysis approach has been deemed appropriate when conflicting or varying conclusions are found in the literature (Cooper, 2017; Tamim et al., 2011).

Meta-analyses assessing the overall effectiveness of technology use on student achievement are ubiquitous in educational literature, with relatively fewer in statistics education. For example, in their second-order meta-analysis investigating bias in meta-analysis studies assessing the effectiveness of technology integration in higher education, Bernard, Borokhovski, Schmid, and Tamim (2014) found that out of the 13 studies collected, four were in statistics education. Furthermore, according to Tamim et al. (2011), over 60 meta-analysis had been conducted since the late 1960's on this topic. Spanning across disciplinary areas, these studies have focused on either the use of one specific technological tool (e.g. statistical software, appellate, simulations, tutorial systems, action response systems (ARS), online/distance education, etc.) (Bernard et al., 2009; Castillo-Manzano, Castro-Nuño, López-Valpuesta, Sanz-Díaz, & Yñiguez, 2016; Means, Toyama, Murphy, & Baki, 2013; Schmid et al., 2014; Sorgenfrei & Smolnik, 2016) or the use of a variety of technological tools (Hsu, 2003; Roh & Park, 2010; Schenker, 2007; Sosa et al., 2011; Tamim et al., 2011) on student achievement. Various meta-analysis approaches such as mixed-effects (Sosa et al., 2011), hierarchical linear modeling (Schenker, 2007), second-order meta-analysis (Bernard et al., 2014; Tamim et al., 2011) have been used. Additionally, some meta-analyses have focused on the effect of specific pedagogical approaches (e.g. cooperative learning, collaborative learning),

small group or individual learning (Kalaian & Kasim, 2014; Lou, Abrami, & D'Apollonia, 2001), student interaction in designed vs contextual treatments (Borokhovski et al., 2016) when using technology.

A long history of primary research examining the effectiveness of using technology to support teaching and learning in post-secondary education compared to not using technology have presented different conclusions on its effects on achievement outcomes. This has led to the increasing presence of meta-analysis research on the phenomenon. The following review of meta-analyses consists of studies conducted in general education and statistics education. Furthermore, reported findings from moderator analyses are organized and discussed according to identified features associated with the design of instruction.

Overall average effect size. These studies have generally reported effect sizes favoring technology-enhanced instruction compared to traditional/non-technology-enhanced instruction on student achievement (Kulik, Kulik, & Shwalb, 1986; Tamim et al., 2011) Effects sizes have ranged in sizes, from small to medium.. Effects sizes have ranged in sizes, from medium. For example, in their study examining the effectiveness of computer-based education (CBE) in adult-education, Kulik, Kulik, and Shwalb (1986) reported a significant overall effect of 0.42 on student achievement favoring CBE. Similarly, Vo, Zhu, and Diep (2017) reported an effect size of 0.38 in favor of blended learning compared to traditional classroom instruction on students' final course grade (achievement). In contrast, in their meta-analysis of 879 studies comparing the effects of technology use in post-secondary classrooms on student achievement, Schmid et al. (2014) computed a smaller overall effect size on achievement of 0.27. Furthermore,

Tamim et al., (2011) reported an overall positive effect size of 0.35 favoring technology use treatment conditions over traditional/non-technology use control conditions. Similarly, in studies examining the effectiveness of technology use specifically in statistics education on student achievement, effect sizes have ranged from 0.24 – 0.57 (Hsu, 2003; 2014; Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011).

Moderator analyses. One of the key components of meta-analytic research is the identification of study characteristics that serve as moderators to explain differences in the estimated mean effects observed across primary studies. Current meta-analyses in general education and statistics education have investigated a diversity and variety of study characteristics, as is often a point of interest in meta-analytic approaches. These have included for example, the coding of substantive study and methodological aspects of the phenomenon under study related to contextual factors (e.g. subject, disciplinary area, student grade level, sample size), modes of instructional delivery, types of technology used, technology features, pedagogical approaches, pedagogical interactions, confounding factors (e.g. teacher bias, treatment/control implementation), research design, publication bias, etc. Additionally, extrinsic characteristics not related to the phenomenon such as publication type and status, etc. have been examined.

Context. Learner characteristics and contextual elements of the instructional environment have been said to be important considerations when assessing the effectiveness of technology-enhanced learning environments (Cobb, 1992). In Tamim et al.'s (2011) second-order meta-analysis on the impact of technology on learning, no significant effects were found for moderators related to subject matter. However, grouping subjects by STEM and non-STEM, Schmid et al (2014) found that in the

technology vs. a no technology use control condition, STEM subjects performed statistically higher than non-STEM subjects (e.g. humanities, education, and language). The opposite was found when comparing the groups across conditions of varying levels of technology use in both treatment and control conditions (non-STEM subjects had significantly higher effect sizes). Given this finding, the authors suggested that further research should investigate the pedagogical approaches that might explain the cognitive underperformance of STEM students. Meanwhile, Vo, Zhu, and Diep, (2017) also reported statistically significant greater effects on student achievement for STEM courses compared to non-STEM. As it relates to student grade level, several studies have found a larger significant effect of technology use on student achievement for studies that sampled graduate students compared to undergraduates (Schenker, 2007; Schmid et al., 2014; Sosa et al., 2011). Sosa et al. (2011) suggests that greater effects for graduate students could be associated with findings that report higher levels of self-regulation and positive attitudes toward statistics for this group compared to undergraduates.

Mode of instructional delivery. Meta-analyses have examined instructional delivery modes as a treatment condition, as well as a moderator. Studies assessing the effectiveness of blended learning on student achievement have produced small to medium size effects favoring blended learning environments (Bernard, Borokhovski, Schmid, Tamim, & Abrami, 2014; Vo, Zhu, & Diep, 2017). However, the content and context in which learning occurs directly influences the blended learning experience (Bednar, Cunningham, Duffy, & Perry, 1992). In their meta-analysis assessing the effectiveness of technology use in post-secondary classrooms, Schmid et al. (2014) examined the influence of blended/classroom instruction on student achievement as a

contextual moderator variable. They reported statistically significant positive effects favoring blended contexts ($g+ = 0.33$) when compared to classroom instruction as a control. As it relates to online learning, Larwin and Larwin (2011) found that the use of technology with face-to-face instruction had the greatest influence on student achievement ($d = 0.539$) with a negative effect size reported for courses delivered online. Meanwhile, Schenker (2007) found no significant differences for studies using online compared to traditional learning formats.

Technology (type, design, function, timing of content presentation). A variety of technologies have been assessed for their moderating effects on student acheivement. Steenbergen-Hu and Cooper (2014) conducted a meta-analysis assessing the effectiveness of various types of intelligent tutoring systems (ITS) (e.g. AutoTutor, Assessment and Learning in Knowledge Spaces, eXtended Tutor-Expert System, and Web Interface for Statistics Education) in post-secondary education and reported an overall positive effect of ITS use on academic achievement ($g+ = 0.32$ to 0.37). The effect of ITS use was found to be greater than traditional clasroom instruction and other pedagogical approaches. Furthermore, as it relates to the timing of instruction, Larwin and Larwin (2011) and Sosa, Berger, Saw, and Mary (2011) found that longer periods of instructional time using technology resulted in statistically significantly higher effects on acheivement.

As it relates to the (pedagogical) function of technology, the use of technology to present information or supplement information has been often associated with higher effect sizes on student achievement. Bernard et al. (2014) found significant effects for technology tools that provided cognitive support in BL environments compared to those

used for supporting/presenting content. Similarly, Schmid et al (2014) and Schenker (2007) found significant effects for technology tools that provided cognitive support compared to those used for supporting/presenting content. Meanwhile, Larwin and Larwin (2011) and Sosa et al. (2011) reported significantly positive effects on student achievement for tools that supplement teaching compared to tools used alone to deliver instruction (e.g. pure online instruction) and for those used for face-to-face instruction. This was mirrored in Tamim et al.'s (2011) second-order meta-analysis, in which a greater significant effect was found for technology use to support instruction (e.g. CAI and CBI) compared to those used to direct instruction (e.g. word processors, simulations). Additionally, in Schenker's (2001) study, statistical analysis software, enhanced lecture, and web-based and online learning were significantly negatively related to effects of technology use on student achievement. Thus, student achievement was lower when using these tools. However, studies using drill and practice produced a significant positive effect size on student achievement (Schenker, 2007). These findings were also reported by Hsu (2003). However, examining differences between teacher produced and commercial tools, Hsu (2003) reported that teacher-made programs were more effective than commercial programs.

Interactions(technology, pedagogy, content). It has been argued that technology alone does not influence learning outcomes **(**Clark, 1994**).** Through moderator analyses, a variety of pedagogical related variables have been examined for their influence on effect sizes observed in technology-enhanced learning environments. In their meta-analysis investigating the effects of social contexts when students use technology to learn, Lou et al. (2001) reported significant average effects for small group learning compared to

individual learning (0.15) and tasks performed in groups (0.31). Additionally, in their study assessing the effectiveness of computer-assisted statistics instruction, Sosa et. al (2011) examined moderators related to the level of learner engagement (e.g. extent of cognitive/active learning) and learner control (e.g. instructor dependent, learner dependent, beyond instructor and learner) and did not report any significant effects on student achievement.

Furthermore, few meta-analyses have examined the effects of the interactions among pedagogical factors in technology-enhanced classroom learning environment. One of these was Bernard et al. (2014) who examined the effects of the interactions among student-student, student-teacher, and student-content on student achievement. Using a sub-collection of experimental studies comparing blended learning (BL) environments and classroom instruction, they reported a significant effect on student achievement (g = 0.334, $p > .01$, k =117) in favor of BL and for the interactions. As it relates to the interaction between technology use and content, Sosa et al. (2011) examined the complexity of statistical concepts presented (e.g. inferential/hypothesis testing to descriptive information), as well as the breadth of concepts covered (e.g. one or multiple topics) when using technology. Although not statistically significant, the authors found larger correlations between technology use and student achievement when more complex concepts were covered.

Assessment. Cognitive outcome measures retrieved from primary studies for the computation of effect sizes have included those measuring student achievement when using technology have included course grades, exams, quizzes, standardized test scores, homework, and assignments, and achievement scales/instruments (e.g. CAOS (Garfield

et al., 2012), ARTIST (Mcgowan & Gunderson, 2010)). Zieffler et al. (2008)warned that studies using these types of measures often fail to report the psychometric properties of the measurement instruments used, resulting in findings that cannot be generalized beyond the studies' context.

In Larwin and Larwin's (2011) meta-analysis, exams and quizzes were the most used outcome measures and were also associated with medium and large effect sizes, respectively. The immediacy of assessment with multiple quizzes was postulated to be a reason for the observed larger effect. This was similar to Sosa et al.'s (2011) findings of greater effects for tools that provided more rapid feedback. Additionally, Sosa examined the effects of embedded assessments and the nature of feedback (e.g. targeted feedback, immediacy of feedback) provided by the tool. They found significantly larger effect sizes on achievement for studies using embedded assessment ($CI_{.95} = 0.36 \leq \mu \leq 0.99$) compared to no assessment ($CI_{.95} = 0.12 \leq \mu \leq 0.40$), and no significant effect for the nature of feedback.

Report characteristics. Report characteristics that have been examined to assess their influence on meta-analysis findings have included publication year, type/source, research design. Some meta-analyses have reported significantly larger effects of technology use for recently published studies (Larwin & Larwin, 2011; Schenker, 2007; Means, Toyama, Murphy, Bakia, & Jones, 2009), while Schmid et al. (2011) found no change over the years they examined. Means attributed higher effect sizes for more recent studies to advancements in technological innovation. Furthermore, Larwin and Larwin (2011) found significant effects for source of research studies (publication type), while Tamim et al. (2011) reported a non-significant effect of publication type (source).

Additionally, Tamim et al. (2011) found no significant differences in effect sizes for research design.

Methodological characteristics. Current technology effectiveness meta-analyses have examined a variety of methodological characteristics to examine factors related to potential bias and confounds associated with the implementation of the treatment.

Publication bias. Publication bias (also known as "File Drawer" problem) is a concern in meta-analysis research whereby studies where significant outcomes are reported are most likely to be published than those reporting non-significant findings (Card, 2012). Testing this, some studies have reported statistically significant larger effects for published studies compared to unpublished (Kulik & Kulik, 1991; Kulik et al., 1980; Schenker, 2007) an indication of publication bias.

Confounds. Clark (1985) challenged reports that concluded significant gains in achievement using computer-based instruction (CBI) when compared to traditional instruction. Through his meta-analysis of a sample of studies examined by Kulik et al. (1980), Clark argued that CBI studies were confounded (effects overestimated) by the instructional methods used in the CBI treatements. He argued that CBI treatment conditions often employ greater efforts in instructional design and development than do the comparison non-CBI control condition. Clark (1985) noted that studies comparing technology use to other media or traditional teaching conditions should employ the same instructional methods in both conditions to teach the same content to avoid confounding effects.

Additionally, same-teacher effects have been identified as potential confounds. Studies that used designs where different instructors taught treatment (technology-

enhanced) and control (no technology use) classes, had significant effects on outcomes measured; whereas non-significance were reported when one instructor administered both treatment and control conditions (Clark, 1985; Kulik et al., 1980). This has been attributed to teacher's unique approaches to designing instruction (Kulik et al., 1980). Teachers may experience a compensentory rivalry effect where either concsiously or unconsiously they mask the true effect of technology-enhanced instruction when sensing a job threat, leading to the underestimation of the true effect (Clark, 1985). Furthermore, Clark (1985) addressed concerns about a novelty effect associated with the length of instruction time where studies in which instruction was carried out in a short term produce greater effect sizes than longer-term studies. This was also found to be the case in meta-analyses conducted by Kulik and Kulik (1991) and Sosa et al. (2011).

Current state of technology effectiveness meta-analysis research

Generally, the body of educational research has adopted the view of technology use as a positive influence on student learning (Archer et al., 2014; Lakhana, 2014). However, substantiating the true effectiveness of technology use on learning through evidence-based research has been a concerned raised by researchers in the field (Roblyer, 2005; Schrum et al., 2007). This has been a result of fragmented findings contained in the literature about its effectiveness on learning. These inconsistencies have further raised discussions and questions about the quality of the studies and the practical usefulness of their findings, both in general educational research (Roblyer, 2005; Ronau et al., 2008) and in statistics education research (Garfield et al., 2008; Zieffler et al., 2008; Hassad, 2014). Current studies have faced criticism for conclusions that attribute positive learning outcomes to the use of specific tools (Schrum et al., 2007). Among these criticisms is the

claim that technology effectiveness studies are confounded by instructional design and teacher-related effects (Clark, 2001; Clark, 1994; Roblyer, 2005). Additionally, in regard to current technology effectiveness research, Archer et al. (2014) raised concern about the lack of attention placed on evaluating implementation fidelity (IF) in primary studies examined; This is despite agreement about IF's potential to significantly impact outcomes measured (Archer et al., 2014; Tamim et al., 2011). Implementation fidelity is concerned with differences in the way technological-based interventions were implemented (Archer et al., 2014). According to Archer et al. (2014), IF is influenced by training and support provided to teachers, teachers' content and technological knowledge, and implementation of intervention by teacher or researcher. Furthermore, identifying the impact of IF and methods for determining its impact on introductory college-level statistical cognitive outcomes are among research priorities recommended by Pearl et al. (2012).

It has been suggested that technology effectiveness research should focus on aspects of instructional design instead of features of technology (Roblyer, 2005, cited in Kozma, 1991) however, research lacks in this area (Roblyer, 2005). Kennewell (2001) asserted that the effectiveness of ICT use in the classroom is contingent upon a variety of factors that should be assessed, including the classroom setting/culture, the pedagogical approaches used, learning tasks and activities, resource availability, how ICT is employed and its purpose, student's perceptions and technological skill. Additionally, the National Technology Leadership Coalition (NTLC) (a cross-disciplinary group of professional members) has emphasized that research on student learning should include considerations of affordances that are provided by technology, pedagogy, and content (Thompson et al.,

2008).Yet, given the complex nature of the learning environment, fewer considerations have been placed on the interrelations between pedagogical strategies, the design of instruction, and content-related features of primary studies that contribute to the impact on learning outcomes (Roblyer, 2005; Spencer, Merrill, Elen, & Bishop, 2014). Bates (2015) argued that the type of technology and the method of delivering instruction are related more to the flexibility and accessibility associated with learner characteristics, while pedagogy and the design off instruction influence learning. Additionally, according to Schrum et al. (2007, as cited in Shulman & Clark, 1983) ,

> Research questions and designs that fail to differentiate by the content being studied, the pedagogical strategies employed, and the way that technology interoperates with these variables will probably continue to find that merely using a technology medium is not educationally beneficial. But research that explores how technology interacts with pedagogy and content may disprove Clark's claim that "media do not influence learning under any conditions (p. 445).

The significance of designed instruction and learning was evident in a meta-analysis study conducted by Borokhovski, Bernard, Tamim, Schmid, and Sokolovskayan (2016). In their study, the authors examined the influences of designed and contextual interaction treatments on student achievement when using technology. Designed interaction treatments were associated with the intentional design/planning of instruction that incorporated collaborative learning when using technology; meanwhile, contextual treatments represented the unplanned use of collaborative learning when using technology. The findings of their study revealed higher positive effects for collaborative learning that was planned/designed into instruction compared to unplanned collaborative

learning (e.g., mean (g) = 0.52, k = 25 vs. (g) = 0.11, k = 20; Q_M = 7.91, p < .02). Additionally, in studies employing designed interaction, the use of tools that supported cognitive learning versus communication tools were associated with higher student achievement. Furthermore, according to Ross and Morrison (2014) a "happy medium" is needed between internal and external validity with research that use strong methodologies and inform instructional design and practice.

A general review of the literature in the fields of instructional design and educational research on assessing technology effectiveness recommend and emphasize the importance of using appropriate theories when assessing learning in an environment in which technology is used (Knowles, Holton III, & Swanson, 2015; Lowyck, 2014). Furthermore, Bernard et al. (2014) suggested that a theoretical framework or rationale is needed for identifying relevant characteristics when assessing the effectiveness between two treatments (intervention and control). However, the field of research lacks a common theoretical framework from which to guide meta-analysis research assessing the effectiveness of technology use. One known example of employing a theoretical framework in meta-analysis is a study conducted by Means, Toyama, Murphy, Bakia, and Jones (2009) on the effectiveness of online learning in all levels of education. The authors developed a conceptual framework to examine evidence-based practices in online learning by identifying three major components related to the type of activity involved. These included 1) the objective for using technology – either as a replacement or enhancement to traditional face-to-face instruction; 2) the pedagogical approach used to elicit a type of learning experience (e.g. expository instruction (receiver), active learning (doer), interactive learning (contributor) based on the extent of control the learner has on

the content and learning activity; and 3) and the synchronous (real-time learning) or asynchronous nature (time lag in instructional content presentation and student response) of communication. Additionally, Bernard et al., (2009) and Borokhovski et al. (2016) used a theoretical framework to guide their selection of substantive characteristics that were related to interaction treatments in distance learning and technology supported environments, respectively. Furthermore, emphasizing the importance of judging the research quality of primary studies in a quantitative synthesis, Cooper and Hedges (1994) commented that "Theoretical considerations are obviously "relevant" to the proper conduct of research synthesis" (p. 100). Similarly, Bernard et al. (2014) suggested that a theoretical framework or rational is needed for identifying relevant characteristics when assessing the effectiveness between two treatments (intervention and control).

In his review of eight meta-analyses conducted in education and social science, Slavin (1995) raised concerns about several methodological weaknesses. The author pointed to issues that included the combining of primary studies measuring related but different outcomes, ignoring the selection bias that may be inherit in primary studies and incorrectly classifying the randomization of sampling units when non-random sampling was actually employed in primary studies. Slavin further argued that these weaknesses can lend to misleading conclusions made. He criticized the traditional meta-analysis practice of exhaustive inclusion of primary studies meeting broad standards in their selection of independent and dependent variables. Additionally, he claimed that meta-analyses often did not incorporate judgements about the quality of the studies selected.

Indeed, despite increasing meta-analysis inquiry in statistics education, the quality of existing research evidence is still in question (Hassad, 2014; Tishkovskaya &

Lancaster, 2012; Zieffler et al., 2011). More information is needed about the reliability, validity, and generalizability of existing evidence, as well as reference to reform initiatives in statistics education (Hassad, 2014). Meta-analysis studies have attempted to provide an overall estimate of the effectiveness of technology use in statistics education, also allowing for findings that are generalizable across contexts (Zieffler et al., 2011). However, despite reporting significant findings, Sosa et al. (2011) concluded that the unexplained residual variance provided an indication that additional study characteristics (moderator variables) may provide meaningful explanations for remaining unexplained differences in the effectiveness of technology use on achievement in statistics. In their study, the authors examined the effects of the complexity of statistical concepts, degree of simulation, and the breadth/range of statistical topics, however, no significant findings were found. This may be partially due to the lack of consideration of the interactions that are a result of the synergies between technology, pedagogy, and content as proposed by Moore (1997). For example, interactions may occur between complexity/breadth of statistical content and pedagogical/technology type (e.g. simulation). Furthermore, previous meta-analyses in statistics education lack a focus on measuring the effectiveness of technology use from a perspective of informing reform-based initiatives. Hassad (2009) argues that the dearth of evidence in literature concerning the effectiveness of reform-based practices in statistics education is a hindrance to the development of the field.

Theoretical frameworks

Recommendations for pedagogical technology-based instruction by statistics educators have emphasized the application of learning theories and principles of

instructional design to support students' understanding of statistical concepts, recognizing that various factors influence effective instruction (2011; Cobb & McClain, 2004; Prodromou, 2015; Tu & Snyder, 2017).

Instructional design

Concerning meta-analysis research, some scholars have urged that a detailed analysis of the characteristics of the learning environment, as described in the context of primary studies should be employed (Kennewell, 2001; Schrum et al., 2007). Given this, Instructional Design (ID) models provide a conceptual framework which outline elements of the design of instruction (Gustafson & Branch, 2002). They serve as a guide for identifying the instructional activities and contexts in which technologies can most effectively be implemented to support learning in face-to-face and online environments (Bates, 2015; Cobb & McClain, 2001; Khalil & Elkhider, 2016). These include considerations of learners' needs, specification of instructional goals, decisions about instructional materials/resources, and the assessment of teaching and learning activities that lead to the achievement of established learning goals and objectives (Bates, 2015; Khalil & Elkhider, 2016).

The early beginnings of the field of instructional design can be traced back to the 1960s - 1970s (Gustafson & Branch, 2002; Khalil & Elkhider, 2016). ID provides a systematic, step-by-step process for designing, implementing, and evaluating planned instruction (Gustafson and Branch, 2007). The principles that are the foundation of ID are based on learning theories associated with behaviorist, cognitivist, or constructivist viewpoints (Gagne, Wager, Golas, Keller, & Russell, 2005; Khalil & Elkhider, 2016; Ozdilek & Robeck, 2009). These learning theories inform ID as they describe the

process, situations and ultimately, the conditions of learning that lead to behavioral changes (Gagne et al., 2005). As a whole, they emphasize learner engagement and the use of educational technologies to foster student learning and achievement, also recognizing that successful pedagogy requires a systematic approach to planning, developing, and executing instruction (Bates, 2015).

Various ID models exist, with the ADDIE model arguably being the most recognized (Göksu, Özcan, Çakir, & Göktas, 2017). As the first ID model to be developed, the ADDIE model has become a general framework used in education and industry from which other models have been developed (Göksu et al., 2017; Gustafson & Branch, 2002). It comprises four key phases of Analysis, Design, Development, Implementation, and Evaluation (Göksu, Özcan, Çakir, & Göktas, 2017; Gustafson & Branch, 2002; Khalil & Elkhider, 2016). These phases have been adopted and modified into other ID models to fit the individual learning context and environment (e.g. constructivist/technology-enhanced) (Hassad, 2011). This has been amid criticisms of the ADDIE model's behaviorist origin that renders it inapplicable to non-traditional learning environments (Bates, 2015; Khalil & Elkhider, 2016).

According to Ozdilek and Robeck (2009), in the analysis phase, learner characteristics and their needs are assessed, gaps in learning are identified and desired learning outcomes (goals of instruction) are established. During the design phase, measurable learning objectives are defined and decisions about the instructional delivery mode, learning activities, and learning materials and tools are specified. In the development stage, learning materials and activities are developed or obtained. Following, is the implementation phase where learning activities and materials are

delivered to the group of learners as instruction is carried out within the learning context. The last stage involves the assessment and evaluation of the delivery of instruction and learning, which can be formative, summative or both and revisions are made, as necessary.

TPACK/TPSK

The role of an individual's knowledge in supporting the effective use of technology is the view adopted by proponents of the TPACK framework. The TPACK framework, which is largely supported in teacher education literature, stresses that instructors' pedagogical, technological, and content knowledge are required for the effective use of technologies that lead to the achievement of intended learning outcomes. This has led to a focus on the types of knowledge required by teachers when using technology for teaching and learning. Harris, Mishra, and Koehler (2009) developed the Technological, Pedagogical, and Content Knowledge (TPACK) framework to guide an understanding of the required knowledge. The use of TPACK as a framework, is recommended as a way to conceptualize the aspects of teachers' knowledge that are necessary for effective teaching when using educational technologies (Harris, Mishra, & Koehler (2009). Furthermore, TPACK emphasizes the interdependencies among aspects of teachers' knowledge when using technology, which are broken into: Technological Knowledge

(TK), Pedagogical Knowledge (PK), Content Knowledge (CK), Technological Pedagogical Knowledge (TPK), and Technological Content Knowledge (TCK) (Harris et al., 2009; Mishra et al., 2006). These include knowledge of: "(a) technological content knowledge about how to teach a subject with technology; (b) instructional strategies and

representations; (c) students' thinking with technology; and (d) curriculum materials that integrate technology" (Prodromou, 2015, p. 32).

Though originally developed to guide effective technology integration in mathematics education, the TPACK framework was designed to be extended for use across various subjects. Building upon this framework, Lee and Hollebrands (2008) developed the Technological Pedagogical Statistical Knowledge (TPSK) framework to guide an understanding of the types of teacher knowledge required to improve learners' understanding of statistical concepts (Lee & Hollebrands, 2008; Makar & Sousa, 2014; Prodromou, 2015). According to the TPSK framework, specialized knowledge required by statistics educators includes:

(1) understanding students' learning about statistical ideas with technology, (2) conceiving of how technology tools and representations support statistical thinking, (3) developing instructional strategies to use in statistics lessons with technology, and (4) critically evaluating and using curricula materials for teaching statistical ideas with technology. (Makar & Sousa, 2014, p. 3)

TPSK has been used to identify a variety of instructional activities associated with relevant content matter and technology use that promote statistical learning (Lee & Hollebrands, 2008; Makar & Sousa, 2014; Prodromou, 2015). Table 1 provides an outline of instructional activities associated with teaching statistics as according to the TPSK framework. It is also noted that technologies are used in diverse ways based on teachers' knowledge in these areas, as well as the affordances and constraints presented by the instructional context (Kennewell, 2001; Schrum et al., 2007). Based on research from student learning in technology-enhanced environment, the following content-related

instructional activities have been associated with the types of statistical knowledge required by teachers (Makar & Sousa, 2014):

Table 1

TPSK Content-Related Activities

TPSK	**Instructional Activity/Content**	**Statistical Learning (Examples)**
Statistical Knowledge (SK)	(1) engaging in exploratory data analysis [EDA], (2) focusing on distributions and describing data as an aggregate, (3) coordinating measures of center and variability in distributions; and (4) considering key differences between statistical and mathematical thinking.	• Examine trends in data, residuals, and correlations • Interpretation of models with support for predictions
Technological Statistical Knowledge (TSK)	(1) automating computations and graphs, (2) exploring data with a variety of representations, (3) visualizing abstract concepts, (4) simulating phenomena, and (5) accessing large data sets	• Dynamic visualization effects of an outlier on correlation and least squares regression line • Use of graphs to conceptualize changes in overlaying of statistical measures (e.g. means, regression lines)
Pedagogical Statistical Knowledge (PSK)	(1) planning for group projects and discussions about data, (2) supporting students in making statistical arguments based on appropriate evidence, and (3) considering the contexts used for teaching statistical ideas	• Make decisions and arguments about statistical investigations • Deliberate a variety of arguments through group discussions

Table 1 (continued).

TPSK	Instructional Activity/Content	Statistical Learning (Examples)
Technological Pedagogical Statistical Knowledge (TPSK)	(1) understanding students' learning about statistical ideas with technology, (2) conceiving of how technology tools and representations support statistical thinking, (3) developing instructional strategies to use in statistics lessons with technology, and (4) critically evaluating and using curricula materials for teaching statistical ideas with technology	The design of learning activities using TSK, PSK, and SK in technology-enhanced environments leads to students' improved statistical knowledge

Note: TPSK components aligned with instructional activities and content as outlined in Makar and Sousa (2014).

Constructivism

Constructivism stems from cognitive theory and is based on the idea that students enter the learning environment with prior knowledge and as they engage in active learning experiences, they construct new knowledge through cognitive and meta-cognitive processing, which leads to the achievement of positive (and higher-order) learning outcomes (Cobb & McClain, 2001; Garfield & Ben-zvi, 2007; Tishkovskaya & Lancaster, 2012). Constructivism posits that for effective learning to occur, the learning condition should be one which supports student engagement and active learning (Garfield, 1995; Hassad, 2011; Lowerison et al., 2006).

The main tenants of the reform movement in statistics education are the development of students' conceptual understanding related to statistical thinking and reasoning, changes in content taught, and improvement of instructional strategies from traditional to learner-centered instructional approaches (Hassad, 2011; Tishkovskaya & Lancaster, 2012). Furthermore, the bases of the recommendations for reform-based

technology integration to achieve these learning outcomes have been driven by a constructivist viewpoint (Everson et al., 2008; Tishkovskaya & Lancaster, 2012). Constructivist approaches to teaching and learning in statistics education are associated with the use of technologies to explore statistical concepts, analyze data, foster active learning, and student inquiry (Rossi A Hassad, 2011; van der Merwe & Wilkinson, 2011). Additionally, they include reform-based authentic learning tasks such as projects, group problem solving, lab exercises, discussions, and cooperative and collaborative learning activities (Garfield, & Ben-zvi, 2008; Garfield, & Ben-zvi, 2007; (Rossi A Hassad, 2011; Tishkovskaya & Lancaster, 2012; van der Merwe & Wilkinson, 2011). For example, Kalaian and Kasim (2014) examined the effects of cooperative, collaborative, and inquiry-based learning approaches on statistics students' academic achievement (statistics exam scores) when using technology. Findings favored small group learning approaches with cooperative and collaborative methods having significantly higher positive effects (0.60) on learning compared to inquiry-based learning.

Furthermore, Cobb & McClain (2001) summarize general recommendations (which align with constructivist approaches) for supporting recommended classroom teaching practices in statistics education that include the following:

- Incorporate more data and concepts.
- Rely heavily on real (not merely realistic) data.
- Focus on developing statistical literacy, reasoning, and thinking.
- Wherever possible, automate computations and graphics by relying on technological tools.
- Foster active learning, through various alternatives to lecturing.

- Encourage a broader range of attitudes, including appreciation of the power of statistical processes, chance, randomness, and investigative rigor, and a propensity to become a critical evaluator of statistical claims.
- Use alternative assessment methods to better understand and document student learning. (p. 6)

Conceptual model for assessing effectiveness of technology use

The current study aims to use the ADDIE model as an underlying framework for the development of a conceptual framework to guide the selection of variables of interest for analysis. Table A1 in Appendix A presents the conceptual framework in the context of the phases of instructional design. As educators and researchers seek to enhance learning outcomes in statistics education using technology, leaders in the field have provided recommendations for its effective use and assessment that are grounded in instructional design principles, theories of learning, and constructivist theory. The operationalization of elements of instructional design are identified and operationalized as discussed in education and statistics education literature regarding the use of technology to support learning. These references include: Bates (2015), Chance, Ben-Zvi, Garfield, and Medina (2007); Cobb & McClain (2004), GAISE College Report ASA Revision Committee (2016), Garfield and Ben-zvi (2007), Means et al. (2009), Moore (1997), Harris, Mishra, and Koehler (2009). The diagram in Appendix A (Table A1) illustrates the conceptual framework to provide a contextual understanding of the instructional elements that contribute to effective technology integration in statistics education. Furthermore, the components and elements in the framework are assumed to

be a non-linear representation of the instructional design activities that contribute to effective instruction.

CHAPTER III - METHODOLOGY

The methodology that was used in the current study is meta-analysis (also referred to as research synthesis). According to Glass (1976), a meta-analysis is the statistical analysis of results obtained from a large group of primary studies measuring the same phenomenon, with the intent of integrating the findings. Furthermore, meta-analysis is appropriate for the current study as it is a technique commonly used to explore the common effect of an intervention of interest obtained from different studies, It also seeks to explain variables that moderate the estimated effect (Borenstein et al., 2017; Cooper & Hedges, 1994; Field & Gillett, 2010).

Furthermore, the analysis of research questions involved the comparison of conditions in which technology is used (treatment) vs. not used (control). Thus, going forward, this is the case described when referring to "the effect of using technology." The study characteristics (moderators) examined were those coded from primary studies that are associated with different elements of classroom instructional design (based on the previously described ADDIE Model). The "implementation elements" are associated with study characteristics related to the implementation phase of instructional design. They represent the synergies between technology, pedagogy, and content and include: learning task (pedagogy and content), scaffolding (technology and pedagogy), and technology function with concept (technology and content).

Study characteristics comprised 24 individual attributes associated with different phases of instructional design which included: academic level, learner's academic background, disciplinary area, location, student gender composition, course, learning goal, learning goal function of technology, content, treatment duration, mode of

instructional delivery, technology design, technology type**,** cognitive outcome function of technology, technology, learning task, learner engagement, learner control, feedback type, specificity of feedback, formative assessment, summative evaluation type.

Report and methodological characteristics of the primary studies were associated with: publication type, publication status, publication source, funded status, publication year, and description of instructional design process, research design, respectively. Finally, quality of study was represented by composite scores derived from an evaluation of the extent of risk of bias (Low, Unclear, High) based on validity attributes related to internal, external, implementation, construct, and statistical conclusion validity characteristics.

As such, the following research questions guided the methodological approach:

Research questions

1. What is the overall magnitude of the effect of using technology on statistics achievement?
    a. Are there statistically significant variations in the estimated mean effects of using technology on statistics achievement across studies?
2. To what extent do 24 study characteristics associated with phases of instructional design moderate the effect of using technology on statistics achievement?
3. To what extent are implementation phase elements associated with interrelations between technology, pedagogy, and content predictors of the effect of using technology on statistics achievement?
4. To what extent do report or methodological characteristics of primary studies moderate the effect of technology use on statistics achievement?

5. To what extent is the quality of primary studies a moderator of the effect of using technology on statistics achievement?

A systematic synthesis of the literature and meta-analysis was guided from the recommendations of Cooper and Hedges (1994), Cooper (2017), as well as Slavin (1995) in his call for best-evidence approaches to traditional meta-analysis. Slavin's (1995) best-evidence approach seeks to add to the rigor of traditional meta-analysis by emphasizing a critical evaluation of the substantive relevance and methodological quality of selected studies. Furthermore, according to Cooper and Hedges (1994), the procedure for conducting a research synthesis involves five stages: 1) problem formulation, 2) data collection, 3) data evaluation 4) data analysis and interpretation, and 5) public presentation.

Problem formation

A synthesis of literature on the effectiveness of technology use in post-secondary introductory statistics education was the basis for the formulation of a problem as presented earlier. Thus, the focus of the study was to assess the impact that the synergies between technology use, pedagogical strategies, and content covered have on students' statistics achievement.

Data collection

Retrieval of studies

The search for relevant studies consisted of the use of various keywords and descriptors coupled with the keyword *"statistics",* The keywords and descriptors that were used to search for relevant studies included: *Keywords* = (technology or computer or computer mediated communication or information communication technology or ICT

or CMC or simulation or multimedia or software or online or computer-based or computer-assisted or distance learning, or distance education or web instruction or tutorial or internet or applet) AND (achievement or learning or cognitive or statistical thinking or statistical reasoning or statistical literacy or effectiveness or evaluation or assessment or performance) AND *Descriptors* = (introductory or post-secondary or tertiary). For example, the first search consisted of *statistics and technology and achievement and introductory*.

Source of studies

The selection of studies was limited to those written in English. Various electronic sources were used to retrieve relevant studies and include journals, reports, dissertations, and conference proceedings to locate published and unpublished primary studies. These included database searches such as: Academic Premier, PsychInfo, EBSCO, ProQuest Dissertations and Theses (PQDT), JSTOR, Education Source, OpenDissertations, Educational Resources Information Circuit (ERIC), and Google Scholar, books/book chapters. Finally, the reference section of selected articles and meta-analysis studies (Hsu, 2003; Larwin & Larwin, 2011; Schenker, 2007; Sosa, Berger, Saw, & Mary, 2011) was searched.

Inclusion and exclusion criteria. Critics have argued that meta-analysis techniques compare studies that vary in the methodological approach, operationalization of variables, measurement approaches, analyses (Borenstein et al., 2009; Kock, 2009; Rosenthal & DiMatteo, 2001; Cooper, 2017) and quality (garbage-in-garbage-out criticism) quality (Rosenthal & DiMatteo, 2001; Cooper, 2017).

To enhance the relevance of the studies selected, as well as the construct and external validity, the inclusion and exclusion criteria was established using Campbell's validity framework, outlining cause, effect(s), participants, time period, and location (Cooper and Hedges, 1994). A study was included if it examined the effect of technology use (construct of cause) on the achievement of cognitive (statistical) learning outcomes (construct of effect), among students in introductory statistics courses (participants), between 1997 and 2017 (time period), in a post-secondary classroom (located internationally). Technologies include tools or software that are used to support teaching and learning of statistical concepts/content. These include technology-based tools related to instructional delivery, data analysis, computing, graphing, simulation, multimedia, Internet.

Additionally, methodological criteria restricted studies to those that used at least one objective criterion for assessing statistics achievements (learning outcome) (e.g., grades, assessment test, etc.), employed an experimental or quasi-experimental (e.g., treatment and control group or two-group pre-post research design), as well as reported relevant statistics for computing a common effect, such as Cohen's *d* effect size (mean, standard deviation (*SD*). Furthermore, the treatment condition involved the use of technology and the control condition did not involve the use of technology. Studies that did not meet the inclusion criteria were excluded. Studies were excluded if they were published before 1997 or after 2018; did not measure technology use associated with classroom learning, used one treatment group and no control group or a control using technology;; assessed outcomes at pre-K – 12 grades or in an intermediate or advanced-

level statistics course; used a cross-sectional or correlational research design; or measured achievement using self-reported measures.

Data evaluation

Coding of studies

The primary studies were coded for report, study, and methodological characteristics (including study quality criteria) which served as descriptive or independent variables, while the effect sizes (Hedges *g*) served as the dependent variable in the meta-analysis. Furthermore, mostly low-inference coding (information provided in research report) with fewer high-inference coding (coder inferred) was used (Cooper, 2017). Coder inferred are those characteristics that are not explicitly presented in the study, however, require the coder's judgment/interpretation of their presence. For example, the extent that studies provided a detailed description of their instructional design were high-inference and were coded as 1) described with roughly enough detail to replicate or 2) described with limited detail

Two raters (the researcher and a trained Ph.D. graduate with major in Educational, Research, Evaluation and Statistics) individually coded the studies to be included in the meta-analysis. Inter-rater reliability was computed using Cohen's Kappa (*K*) to assess the level of agreement between coders. The Kappa statistic and percent agreement was reported. The following was used to interpret the IRR index of agreement: less than 0.4 = poor; 0.40 – 0.59 = fair; 0.60 – 0.74 = good; 0.75 and greater = excellent (Cooper & Hedges, 1994). In the case where $K < 0.4$, those studies with low inter-rater agreement were further deliberated to first gain consensus between the raters, and then by

seeking out the consultation of an independent third-party if needed. In the case of continual non-agreement, the study was discarded from the analysis.

Study characteristics. Report and study (instructional elements) characteristics were coded. Report characteristics included (a-d): (a) publication type, (b) publication source, (c) funded research, (d) publication year. The instructional elements characteristics included (e-ab): (e) academic level, (f) learners' disciplinary background, (g) course disciplinary area, (h) location, (i) student gender composition, (j) course name, (k) learning goal, (l) learning goal of technology use, (m) cognitive outcome function of technology, (n) content, (o) treatment duration, (p) content/topic, (q) instructional delivery mode, (r) technology type, (s) technology design, (s) learning task, (u) learner engagement, (v) learner control, (w) scaffolding, (x) feedback type, (y) technology function with concept, (z) formative assessment measure, (aa) summative evaluation type, (ab) summative evaluation measure.

Each of the 24 study characteristics align with a phase of the ADDIE instructional design model, as well as further describe the instructional context, content, and interrelations between technology, pedagogy, and content as presented in the articles. For example, the Analyze stage includes elements that relate to assessing learners and identifying what is to be learned (e-n); the Design phase includes elements that relate to how content is to be learned (o-p); the Develop phase includes elements that relate to production and/or acquisition of instructional materials (q-r); the Implement phase includes elements that relate to the use of material and pedagogical strategies to deliver instruction (s-y); and the Evaluation phase includes elements that relate to monitoring and

assessing the effectiveness of instruction (z-ab). Table A1 in Appendix A provides an outline of the coded characteristics and their operationalizations.

Methodological characteristics. Methodological characteristics were comprised of attributes related to design and implementation features of primary studies. Furthermore, these are operationalized below and included (ac-ae): (ac) material equivalence, (ad) research design, and (ae) description of instructional design process.

Study quality. Study quality included an evaluation of evidence addressing six concerns of threats of validity across primary studies (e.g. internal, external, implementation fidelity, construct, statistical conclusion). These were further evaluated in relation to the extent of risk of bias present (e.g. low, unclear, high).

Operationalization of variables of interest

Operationalization of variables was based on a review of the literature and are described as follows:

Grade level. The grade level of the statistics course being taught was coded according to their undergraduate or graduate level status.

Disciplinary area. The disciplinary area was coded based on the disciplinary field in which the statistics course is taught. These were coded according to the following categories: interdisciplinary (mixture of disciplines), discipline focused (e.g. nursing, math, business, biostatistics, etc.).

Learning goal. According to Garfield, Chance, Poly-San, and Obispo (1999), the reform-oriented learning goals for students learning statistics are: understand the purpose and logic of statistical investigations, understand the process of statistical investigations,

learn statistical skills, understand probability and chance, develop statistical literacy, develop useful statistical dispositions, develop statistical reasoning.

Learning goal function of technology. The effectiveness and usefulness of technology in supporting students' understanding of statistics is driven by the usefulness in supporting learning goals through its of its functional capabilities associated with: automation of calculations, collaboration and student involvement, investigation of real-life problems, simulation used as teaching tool, visualization of concepts, multiple (Garfield et al., 2008).

Content/topic. Content relates to main topics generally covered in introductory level statistics education and which participants are to learn while using technology. Topics commonly taught in introductory statistics courses include: descriptive statistics, hypothesis testing, centrality, variability, distributions, probability (chance & uncertainty), randomness, sampling, inferential statistics.

Mode of instructional delivery. The method in which instruction is delivered was categorized as: face-to-face, pure online, hybrid, flipped

Technology/media type. The type of technology used to support learning was categorized as: commercial statistical package, educational data analysis tools, web or computer-based applet/visualization, stand-alone simulation software, web information resource, drill and practice tutorial, screencast tutorial, LMS/CMS, Clicker, other.

Technology design. The method in which technology is acquired was categorized as: institution hosted, propriety (commercial), or instructor/researcher designed.

Cognitive outcome function of technology. The functionality provided by the technological tool was categorized according to behavioral, cognitive, or constructivist

features that support learning (Spector, Merrill, Elen, & Bishop, 2014). Spector, et al. (2014) described these as: information seeking, information presentation, knowledge organization, knowledge integration, knowledge generation.

Timing of content presentation. The timing presentation of content to learners was categorized as either synchronous (real-time learning - immediate) or asynchronous (time lag).

Learning task. The synergy between content and pedagogy was examined through the types of learning tasks used to deliver content matter to enhance learning. The learning tasks that students engage in was categorized as either assignments or problem solving, laboratory exercises, or multiple.

Learner engagement. The extent at which learners are engaged while using technology was categorized as following: individual, cooperative/collaborative/collective, or mixed.

Learner control. The extent of control that learners have when using technology was categorized as: learner with materials or learner with others.

Scaffolding. The synergy between technology use and pedagogy was captured by the presence or lack of scaffolding provided by the learning tool or instructor and was recorded as: scaffolding present or no scaffolding present.

Feedback type. The extent to which the technology provides feedback was recorded as: feedback or no feedback.

Technology function with concept. The synergy between technology use and content learned was captured by the combination of the functionality of the technology used and the concept learned, and was categorized as: computing (data analysis,

bootstrap)/graphing (distribution, outliers, models, centrality/spread); course management (collaboration); data exploration ; simulation (probability, variability); or multiple.

Formative assessment measure. The types of measures used to monitor students' performance for the purpose of providing feedback to address any misconceptions and guide teaching and learning. These were categorized as: homework assignment/practice questions/activities, tests/quizzes, or multiple measures.

Summative assessment measure. Type of measure used to measure learners' overall learning performance. These were categorized as: another achievement test (e.g. teacher made final exam/test/quiz), standardized achievement/cognitive test, mixed (combined), or both.

Summative evaluation type. The type of assessment measure used to evaluate learners' cognitive performance as it relates to either: authentic assessment (e.g., assignment/project), non-authentic assessment (e.g. course grade/exam/test), or both.

Author. The name(s) of the author(s) was recorded.

Publication year. The year that the article was published was recorded.

Publication status. The publication status of the study was categorized as either published (journal article, book), or unpublished/grey literature (dissertation, MA thesis, private report, government report, conference paper).

Research design. The research designed used was coded according to: independent groups post-test or independent groups pre-test post-test.

Material equivalence. Whether the same or slightly different sets of material were used for the treatment and control group.

Study quality. Study quality is defined as the fit between the primary study's research goal and the characteristics of a study's design and implementation (Cooper, Harris, Hedges, Larry V., Valentine, 2009). A quality scale was used when evaluating design and implementation characteristics of primary studies. Design and implementation encompass elements related to validity concerns as outlined by Shadish, Cook, and Campbell (2002). This also includes issues unique to the fidelity of implementation of technology-based treatments in the classroom (e.g. instructor bias, equivalence of curriculum material). Study quality is also operationalized as the extent of risk of bias (extent that evidence is provided that validity concerns were addressed appropriately or not addressed)

Extent of risk of bias. Study quality is described in relation to the extent of risk bias. This is, the extent that there is evidence that favorable validity attributes have been addressed appropriately (low risk of bias) or not addressed appropriately (high risk of bias).

Developing the study quality scale

The quality of studies included in the meta-analysis were assessed by the researcher using a researcher developed scale that assessed the extent of risk of bias. Additionally, the assessment was conducted to inform gaps in the literature related to evaluation of the methodological soundness of studies. A scale was developed using recommendations related to: 1) validity and reliability concerns in scientific research by (Shadish, Cook, & Campbell, 2002), 2) evaluating risk of bias in systematic reviews (The Cochrane Collaboration, 2011), and 3) the implementation of educational technology as discussed in the educational technology literature. The study quality scale consisted of

five validity attributes that were assessed, which included: construct validity, external validity, internal validity, implementation fidelity validity, and statistical conclusion validity. Studies were assessed based on validity statements related to each validity attribute. For example, when evaluating internal validity, one of the statements asked, "*Was the control group made aware of the treatment condition?*" (Design contamination). Similarly, when evaluating implementation fidelity, one of the statements asked, "*Was the implementation of curriculum the same for both conditions?*" (Equivalence of curriculum material). Table B1 in Appendix B lists the statements that addressed design and methodological threat of validity concerns.

Furthermore, response options for the validity statements were operationalized based on a determination of the extent of risk of bias which is related to whether the validity concern was addressed in each article. The operationalization of response options included the following: the validity concern was explicitly explained and handled correctly; the evidence provided about whether the validity concern was addressed was not sufficient to make a clear determination of extent of risk of bias; and an explanation was provided of how the validity concern was handled but it was handled inappropriately. Each of these response options were associated with a "Risk of Bias Category" (ROB) rating of either low risk of bias, unclear risk of bias, or high risk of bias, respectively. Possible points for risk of bias categories for each statement within a validity attribute ranged from 0 to 2. Points for ROB categories were assigned as follows: "0 points" or Low ROB (evidence was provided but the concern was not handled appropriately); "1 point" or Unclear ROB (insufficient evidence to determine extent of ROB), "2 points" or High ROB (evidence was provided and the concern was handled appropriately). Higher

scores indicated low risk of bias (higher quality – validity concerns addressed), while lower scores indicated high risk of bias (lower quality – validity concerns not addressed), with unclear risk of bias falling in the middle range (insufficient evidence to determine whether validity concern was addressed). Table 2 presents the risk of bias categories and the criterion associated with each (as suggested by The Cochrane Collaboration (2011)), as well as the allotted points.

Table 2

Risk of Bias Categories

Risk of Bias Category	**Criterion**	**Points**
Low Risk of Bias	Explicitly explained in the paper how this risk of bias was handled, and it was handled properly	2
Unclear Risk	There is insufficient information to assess whether an important risk of bias exists; or Insufficient rationale or evidence that an identified problem will introduce bias. The risk of bias is genuinely unknown despite sufficient information about the conduct	1
High Risk	Explained how this risk of bias was handled but it was not handled appropriately	0

Note: Table of risk of bias categories with their associated criteria and allotted points. The table is reproduced from recommendations in Cochrane Handbook, Chapter 8: Assessing risk of bias in included studies (The Cochrane Collaboration , 2011)

Evaluating overall ROB across studies. Each risk of bias attribute was associated with one or multiple validity concern statements. Given this, scale ranges were created for each validity attribute by computing the highest possible total points given the number of statements and dividing it into three segments – representing ranges for low,

unclear, and high risk of bias. For example, the total possible points for "implementation fidelity" was six points (with six points representing three statements in which each provided evidence that the risk of bias concern was explicitly explained and appropriately handled – two points per statement). When dividing the possible points into three ROB categories, scale points were allotted according to the following ranges: "0-2" (High ROB), "3-4" (Unclear ROB) and "4-6" (Low ROB).

Furthermore, total risk of bias scores were computed within individual studies and across studies, segmented by validity attributes. A summary of risk of bias provided an overall breakdown of the proportion of studies in each ROB rating category as a function of the total ROB scores across all categories. The total possible study quality points when combining points across all studies by risk of bias categories ranged from: 0-12 (High ROB), 13-24 (Unclear ROB), 25-36 (Low ROB). When the range of possible points were not evenly divisible by three, wider ranges were allotted to the categories associated with greater extent risk of bias – aiming at a conservative approach to assigning bias. Finally, the interpretation of the summary assessment of risk of bias was guided by recommendations outlined in Table 8.7.a. in the Cochrane Handbook (Chapter 8: Assessing risk of bias, 2011) and is presented in Table 3.

Table 3

Criteria for Summary Assessment of Risk of Bias

Risk of bias	**Interpretation**	**Within a study**	**Across studies**
Low risk of bias	Plausible bias unlikely to seriously alter the results.	Low risk of bias for all key domains.	Most information is from studies at low risk of bias.

Table 3 (continued).

Risk of bias	Interpretation	Within a study	Across studies
Unclear risk of bias	Plausible bias that raises some doubt about the results.	Unclear risk of bias for one or more key domains.	Most information is from studies at low or unclear risk of bias.
High risk of bias.	Plausible bias that seriously weakens confidence in the results.	High risk of bias for one or more key domains.	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.

Note: Summary assessment of risk of bias reproduced from Cochrane Handbook, Chapter 8: Assessing risk of bias in included studies; Table 8.7.a. Retrieved from: https://handbook-5-1.cochrane.org/chapter_8/table_8_7_a_possible_approach_for_summary_assessments_of_the.htm

Calculating effect sizes

Prior to analysis, standardized effect sizes (Hedges' *g*) were computed for each primary study. Standardized mean differences are useful when outcomes are measured differently across studies (Cooper, 2017; Cooper & Hedges, 1994). Hedges *g* standardizes the measured outcomes in group contrasts, allowing for comparisons to be made across groups. Effect sizes were computed for each study using Hedges' (1981) formula for the standardized mean difference for two independent groups:

$$d = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{SD_P} \quad (1)$$

where $\overline{X}_{G1}$ is the mean outcome of group 1 (e.g., treatment group), $\overline{X}_{G2}$ is the mean outcome of group 2 (e.g., control group) SD_p is the pooled within group standard deviation and is computed using the formula:

$$SD_p = \sqrt{\frac{(n_{G1}-1)SD_{G1}^2+(n_{G2}-1)SD_{G2}^2}{n_1+n_2-2}} \quad (2)$$

where n_{G1} and n_{G2} are the sample sizes for group 1 and group 2, respectively, and SD_{G1} and SD_{G2} are the respective group standard deviations. Furthermore, the following

(3)

formula was used to compute the standardized mean difference for studies that used an independent groups pre-test post-test design (Morris & Deshon, 2002):

$$d_{\text{IGPP}} = \frac{\bar{X}_{\text{Post,G1}} - \bar{X}_{\text{Pre,G1}}}{SD_{\text{Pre,G1}}} - \frac{\bar{X}_{\text{Post,G2}} - \bar{X}_{\text{Pre,G2}}}{SD_{\text{pre,G2}}}$$

where $\overline{X}_{\text{Post,G1}}$ and $\overline{X}_{\text{Pre,G1}}$ are the mean post-test and pre-test outcomes of group 1 (e.g., treatment group), respectively; and $\overline{X}_{\text{Post,G2}}$ and $\overline{X}_{\text{Pre,G2}}$ are the mean post-test and pre-test outcomes of group 2 (e.g., control group), respectively; and $SD_{\text{Post,G1}}$ and $SD_{\text{Post,G2}}$ are the pre-test standard deviations for each group. According to Hedges (1981), where studies use small samples size (e.g. less than 20), Cohen's *d* effect size index tends to be biased upwards. Therefore, unbiased (Hedges' *g*) effect size estimates were computed for each study from Cohen's *d* using Hedges' (1981) weighted least squares estimation method. The formula for the unbiased (weighted) effect size (*g*) was:

$$g = [1\text{-}\frac{3}{4\text{N-}9}] \quad (4)$$

where *N* is the total sample size ($n_{G1} + n_{G2}$) and the standard error $SE_{(g)}$ of the corrected effect size is computed as:

$$SE(g) = \frac{n_{G1}+n_{G2}}{n_{G1}n_{G2}} + \frac{(d)^2}{2(n_{G1}+n_{G2})} \text{ (Cooper, 2017)} \quad (5)$$

The confidence interval for Cohen's *d* and Hedges' *g* effect sizes (ES) were computed using the formula:

$$\text{ES-}1.95\sqrt{\text{SE}_{\text{ES}}} \leq \text{ES} \geq \text{ES} + 1.95\sqrt{\text{SE}_{\text{ES}}} \quad (6)$$

When studies did not report the mean and standard deviations to directly compute the standardized mean difference but provided other relevant statistical measures (e.g. *F*, *t*, *p*-value), formulas recommended by Lipsey and Wilson (2001) and Boreinstein et al.

(2017) were used to compute an effect size. In some cases, non-independent (stochastically dependent) outcomes measuring statistics achievement were found within studies. When adjustments are not made for non-independent observations, it leads to an underestimation of the variance Cheung (2019). Therefore, when this dependency occurred, an adjustment was made to compute an aggregate effect size and variance, assuming a correlation of .50 between outcomes within the study.

Data analysis and interpretation

The analyses of research questions were conducted using R (version 3.6.2; R Core Team, 2018), and the metafor package (Version 2.1.0; Viechtbauer, 2010). The programs were used to estimate the common effect size and its significance, to conduct moderator analyses using mixed-effects models, and to conduct diagnostic and outlier analyses and produce plots. The use of a random/mixed effects model opposed to a fixed-effect model is based on the assumption that studies were randomly sampled from a larger population of studies (Hedges and Vevea, 1998). Therefore, there exists a distribution of effect sizes in which variations (heterogeneity) exist among their true effect sizes (Cooper, 2017; Viechtbauer, 2010). This heterogeneity could be attributed to unknown methodological differences such as in research implementation, instrumentation, sample characteristics, setting, etc. (Borenstein, Hedges, Higgins, & Rothstein, 2017; Cooper & Hedges, 1994; Field & Gillett, 2010). Furthermore, the random/mixed effects model allows for findings to be generalized to the larger population of studies already conducted, that could have been, and that will be conducted in the future (Cooper & Hedges, 1994; Rosenthal & DiMatteo, 2001; Viechtbauer, 2010). Whereas, fixed effects models allow for inferences

to be generalized only to the sample used (Cooper & Hedges, 1994; Rosenthal & DiMatteo, 2001; Viechtbauer, 2010).

The following describes decisions made in the data analysis and interpretation of meta-analysis findings. These are related to 1) assumptions, 2) examining diagnostics, and 3) conducting the meta-analysis to answer the research questions.

Assumptions

The following assumptions were made about the distribution of the sample data in order to make inferences about the population: 1) all primary studies measure the same phenomenon, 2) the effect size outcomes from each study are independent of one another, and 3) appropriate methods were employed by the primary researcher in the computation of outcomes for each study (Cooper, 2017). The assumption of normality was examined visually through the inspection of a histogram showing the distribution of the studies' estimated effect sizes, as well as through the computation of pseudo *z* scores to assess skewness and kurtosis.

Diagnostics

Outliers and influential cases. Outliers were examined through the inspection of a forest plot showing the observed outcomes and the pooled estimate based on the random-effects model. Additionally, standardized residuals and Cook's distances were used to examine whether studies may be outliers and/or influential in the context of the random-effects and mixed-effects models. Recommendations from Viechtbauer and Cheung (2010) were used to evaluate outlier and influence diagnostics. Standardized residual is a measure of the difference between the average effect size and the effect size of the *i*th study, divided by the estimated standard deviation. Studies with a standardized residual

larger than ± 1.96 are considered potential outliers. Cook's distance examines changes in the fitted values of *k* studies when the *i*th study is removed. Studies with a Cook's value larger than the median plus six times the interquartile range of the Cook's distances were considered to be influential.

Leave-one out. A "one study removed" analysis was conducted to examine potential outlier cases using the random-effects model. As a study is removed, simultaneously, the average effect size is recalculated, and the leverage effects are examined. Changes in the significance of the effect size when a study is removed would indicate that the study influences the distribution of average effect sizes (Bernard et al., 2014).

Analysis of research questions

Random-effects model

As previously mentioned, random and mixed-effects models were used to address the research questions. A random-effects model was used to address research question one. The estimation of the average mean effect was based on the assumption that the observed effects represent a random sample from a super population of true effect sizes and are unbiased, normally distributed, and with variance known (Cooper & Hedges, 2009; Viechtbauer, 2010). The assumption is that:

$$y_i = \theta i + e_i \quad (7)$$

where y_i represents an estimate of the true effect θi with sampling error e_i such that $e_i \sim N(0, vi)$. The *rma* function in the metafor package was used when fitting the random-effects model and residual variance (heterogeneity) was estimated using restricted maximum likelihood (REML) as it provides an unbiased estimate of

(8)

heterogeneity. Furthermore, the mean parameter (μ) was estimated using weighted least squares with weights equal to:

$$w_{\mathrm{i}} = \frac{1}{v_i + \hat{\tau}^2} \tag{8}$$

where $\hat{\tau}^2$ is an estimate of τ^2. The random-effects model was used to estimate the true mean effect (θ_i) and total variability (heterogeneity/between study-variance) (τ^2) exists across effect sizes. The model is represented by:

$$\theta_i = \mu + u_i \tag{9}$$

where $u_i \sim N(0, \tau^2)$, such that the assumption is that the true effects are normally distributed with a mean μ and variance τ^2 (total amount of heterogeneity). Homogeneity among effect sizes is assumed if $\tau^2 = 0$ (e.g., $\theta_1 = \ldots = \theta_k \equiv \theta$), rendering $\mu = \theta$ as the true effect (Viechtbauer, 2010).

Heterogeneity

A test of homogeneity (Cochran's *Q* - test) (Hedges & Olkin, 1985) tests the null hypothesis that there is no statistically significant variation in effect sizes across studies H_0: $\tau^2 = 0$. The *Q* statistic with k-1 degrees of freedom (df), the corresponding *p* value, and confidence intervals were reported. Homogeneity of the variances in effect sizes is assumed if the *p*-value from computing *Q* is not significant at alpha = .05 level (95% CI). Significance is concluded if $p < .05$, providing an indication that the effect sizes are heterogeneous across studies. Additionally, I^2 provides an indication of the proportion of residual heterogeneity to unexplained variability that remains (intra-class correlation) and H^2 provides a ratio of unaccounted variability to sampling variability (variation to signal

ratio) (Raudenbush, 2009, Viechtbauer, 2010). The *rma* package uses the following equations to compute I^2 and H^2:

$$I^2 = 100\% \times (Q - (k-1))/Q$$

$$H^2 = Q/(k-1) \quad (10)$$

where Q represents the test of the heterogeneity and k, the number of studies.

Higgins et al. (2003) provides the following recommendations for interpreting the amount of heterogeneity (I^2): 0% (no heterogeneity), 25% (low heterogeneity), 50% (moderate heterogeneity), and 75% (high heterogeneity). The presence of heterogeneity in effect sizes provides an indication of the distribution of effect sizes around the population mean. Significant heterogeneity signals the analysis of moderator variable to explain differences in the variations of effect sizes observed across primary studies that are due to beyond sampling error (Field & Gillett, 2010).

Mixed-effects model

Separate mixed-effects models were used to conduct subgroup, moderator, and meta-regression analyses to answer research questions two through five. The mixed-effects model allows the inclusion of moderator variables (study-level) that may attribute to some of the heterogeneity observed in the true effects. This results in an approach to fitting a model that accounts for the fixed-effects (within-study) and random-effects (between-study). As in a traditional Analysis of Variance, variables were included in the model as categorical variables (factors). The factor function in R program (R Core Team, 2019) was used to dummy code the variables, with a "1" signifying the presence of a particular attribute within a category and "0" for non-presence. For example, as it relates to the variable *Technology Design*, studies reporting "Teacher"/researcher developed"

were dummy coded “1”, while all other studies where this category was not present were coded “0”. Furthermore, residual heterogeneity was estimated using REML. The model is represented by:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + u_i \qquad \text{(Viechtbauer, 2010)}$$

where x_{ij} represents the value of the j-th moderator variable for the i-th stuc (12)

the assumption that $u_i \sim N(0, \tau^2)$. τ^2 represents the residual variability (heterogeneity) that exists and thus signifies the need for additional moderators to be included in the model.

Publication bias

Oftentimes referred to as the “file drawer problem,” publication bias is related to bias due to unpublished studies that have not been accounted for in the literature. A reason why studies are not found in the literature might be due to non-significant findings, resulting in potential over-stating of meta-analytic findings (Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). To address publication bias, sensitivity analysis was conducted using mixed-model subgroup analysis to examine if the factor, *Publication Status*, was a moderator of the effect size (Card, 2012; Cooper, 2017). A significant test of moderator ($p < 0.05$) would provide an indication of possible publication bias. Secondly, funnel plots provided a graphical approach for examining publication bias. The funnel plot resembles a scatterplot in which effect sizes (x-axis) are plotted relative to their standard error (y-axis), centered around the estimated average effect (Viechtbauer, 2010). A symmetric (funnel-shaped) distribution of observations provides an indication of no publication bias. Furthermore, Egger’s Test (Egger, Smith, Schneider, & Minder, 1997) was used to provide an additional approach to statistical inference regarding the existence of publication bias. Using a regression approach, a significant finding ($p < .05$) would

indicate a linear relationship between a study's sample size and the size of the effect, suggesting publication bias (Egger, Smith, Schneider, & Minder, 1997).

CHAPTER IV – RESULTS

A five-step process was used to guide the meta-analysis and examine findings. First, a search was conducted for primary studies that used independent groups post-test or pre-test post-test designs with a control group to investigate the effect of a technology intervention on statistics achievement. Second, the inter-rater reliability was computed for the coding of studies characteristics. Third, the descriptive findings of study characteristics of the primary studies were evaluated according to their association with the five phases of the ADDIE model. Fourth, the results of the random-effects model analysis and moderator analyses using mixed-effects models that address the research questions were examined. Fifth, the results of study quality and publication bias analyses were assessed with considerations of the evidence they provide for the robustness of the conducted meta-analysis. The analyses were performed using R program (R Core Team, 2019) with the use of the metafor package (Viechtbauer, 2010).

Retrieval of primary studies

Multiple database sources were used to retrieve primary studies. Table 1C in Appendix C provides a list of the databases and keyword searches used. Keyword database searches and records identified through other sources (online search engine and reference lists of existing meta-analyses) resulted in a random selection of 1,399 studies being located (including duplicates). After duplicates were removed, 149 articles were screened through review of their abstracts. Articles were retained if they met the inclusion criterion of having evaluated technology use in statistics education. The exclusion of studies at the abstract review stage resulted in 86 studies that were further inspected by examining their match with all aspects of the inclusion/exclusion criteria

(Table D1 in Appendix D lists the studies that were excluded with explanations). The search resulted in a final selection of k = 32 primary studies (k represents the number of studies) that met all criteria for inclusion in the meta-analysis as shown in Figure 1.

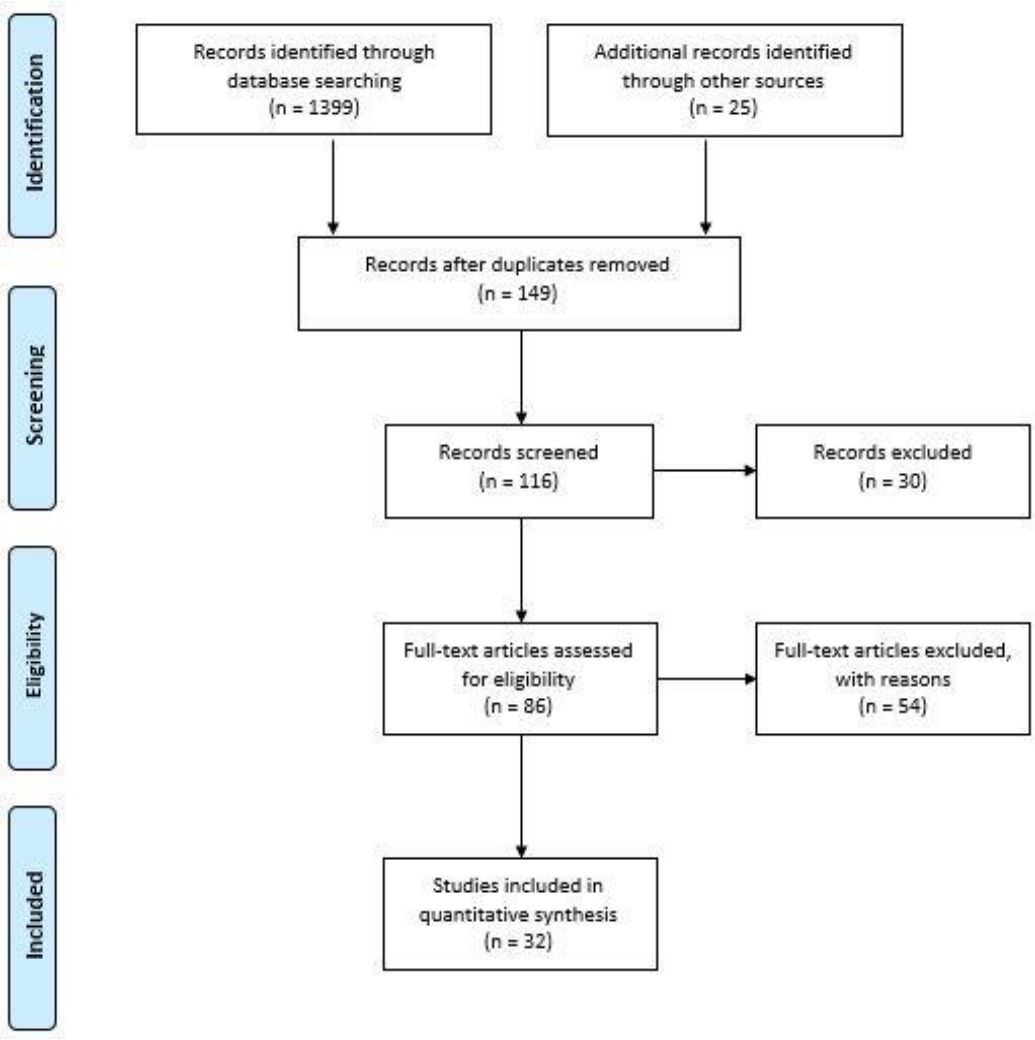

Figure 1. Diagram of Article Selection Process.

Note: Diagram of article selection process modified and adapted from The PRISMA Group, *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement* (Moher, Liberati, Tetzlaff, & Altman, 2009).

A list of these studies is presented in Table 4 with selected coded characteristics related to the inclusion criteria. Overall, the 32 studies reported 42 separate outcomes related to student achievement based on 32 separate samples of students. For example,

Burruss and Furlow (2007) reported four outcomes, separating effects on student's learning of different statistical content areas/literacy (chi-square test, computation, definition, and definition with interpretation). Hilton and Christensen (2002) also reported four outcomes of students' performance on four exams. Furthermore, it was found that there was missing data on five variables across studies. Seven studies did not provide necessary information to categorize the disciplinary background of learners in the classroom as interdisciplinary or same discipline. Similarly, the disciplinary area in which the course was taught could not be determined among seven studies. Only eighteen studies reported information about the geographic location of the institution where classroom instruction occurred. Finally, descriptions of the composition of gender among participants was provided by only 19 studies.

Table 4

Selected Characteristics of Studies Measuring Technology Versus No Technology Conditions

Author (Year)	Academic Level	Research Design	Comparison	*N* Tre	*N* Ctr	*N* Out	*g*	*SE*
Aberson et al. (2003)	U	IP	Tutorial vs. No tutorial	15	10	1	-0.26	.41
Aberson et al. (2000)	U	IGPP	Interactive tutorial vs. Lecture	55	56	1	0.25	.19
Arena & Schwartz (2014)	U	IGPP	Digital game vs. No digital game	14	13	2	0.10	.28

Table 4 (continued).

Author (Year)	Academic Level	Research Design	Comparison	*N* Tre	*N* Ctr	*N* Out	*g*	*SE*
Basturk (2005)	G	IGP	Web information resource vs. Lecture-only	65	140	2	1.10	0.13
Benedict & Anderton (2004)	U	IGP	Jitt vs. Classroom/ traditional	56	67	1	0.39	0.18
Burruss & Furlow (2007)	U	IGP	Visual tutorial vs. Lecture	38	32	4	0.23	0.12
Ciftci, Karadag, & Akdal (2014)	U	IGPP	Commercial stats package vs. Traditional	48	49	1	1.22	0.21
Dinov, Sanchez, & Christou (2008)	U	IGP	SOCR vs. Lecture	88	83	1	0.72	0.15
Frederickson, Reed & Clifford (2005)	G	IGP	Web supported vs. Lecture supported	8	8	1	-0.52	0.47
Gonzalez & Birch (2000)	U	IGP	Computer-based tutorial vs. Traditional/ lecture	29	14	2	0.73	0.28
High (1998)	U	IGP	Computer software vs. Lecture-based	43	44	1	0.26	0.21

Table 4 (continued).

Author (Year)	Academic Level	Research Design	Comparison	*N* Tre	*N* Ctr	*N* Out	*g*	*SE*
*Hilton & Christensen (2002)	U	IGP	Multimedia vs. Overhead transparencies	2801	2801	4	-0.15	0.01
Jones (1999)	U	IGPP	Web-based (online/Internet) vs. Traditional classroom	24	46	1	0.50	0.25
Lane & Aleksic (1998)	U	IGPP	Course website/ lab vs. Lecture-based	1597	340	3	0.42	0.05
Larwin & Larwin (2011)	U	IGPP	Simulation vs. No simulation (traditional)	27	27	2	-0.64	0.20
Lloyd & Robertson, (2012)	U	IGP	Video tutorial (screencast) vs. Text tutorial	26	26	2	-0.54	0.28
Lu & Lemonde (2013)	U	IGP	Online vs. Face-to-face	20	72	1	0.08	0.25
Maurer & Lock (2016)	U	IGPP	Simulation-based vs. Traditional inference curricula	50	51	1	-0.54	0.20
McLaren, (2004)	U	IGP	Online vs. Classroom/ lecture	80	127	1	0.00	0.14

Table 4 (continued).

Author (Year)	Academic Level	Research Design	Comparison	*N* Tre	*N* Ctr	*N* Out	*g*	*SE*
Milic, et al. (2016)	U	IGP	Blended vs. Traditional/ lecture	87	353	1	0.78	0.40
Mills (2004)	U	IGPP	Computer simulated methods vs. Traditional	13	17	2	-0.32	0.27
Morris (2001)	U	IGPP	Computer-based learning system vs. Paper-based	17	16	1	0.96	0.34
Peterson (2016)	U	IGP	Flipped vs. Traditional lecture	24	19	1	0.16	0.31
Petta (1999)	G	IGP	Web-based management system vs. Traditional lecture	11	24	1	0.27	0.36
Ragasa (2008)	U	IGPP	Computer-assisted instruction vs. Traditional method	38	15	1	0.94	0.31
Smith (2017)	U	IGPP	Gamified module vs. No gamified module	24	32	1	1.09	0.27
Spinelli (2001)	U	IGP	Technology (Minitab) vs. Traditional	41	69	1	0.27	.019

Table 4 (continued).

Author (Year)	Academic Level	Research Design	Comparison	*N* Tre	*N* Ctr	*N* Out	*g*	*SE*
Tintle, et al. (2018)	U	IGPP	Early simulation-based inference vs. Consensus	289	366	1	0.53	0.20
Utts, et al. (2003)	U	IGP	Hybrid vs. Traditional	199	76	1	0.30	0.08
Wang & Newlin (2000)	U	IGP	Web- based vs. Face-to-face	49	66	1	0.12	0.28
Wang. (1999)	G	IGPP	Computer vs. Reading only	12	11	2	0.27	0.31
Wilmoth & Wybraniec (1998)	U	IGP	Presentation software vs. No presentation software	125	108	1	0.25	0.13

Note: * Equal sample size (treatment and control) assumed; U = Undergraduate, G = Graduate; IGP = Independent Groups Post-Test, IGPP = Independent groups pre-test post-test; *N* Tre and *N* Ctl = number of outcomes in the treatment and control groups, respectively; *N* Out = number of outcomes.

Inter-rater reliability

All primary studies (k =32) were coded by the author and a second rater (a third-year PhD students) who received training in the coding process. Cohen's Kappa was used to assess the reliability of the coding of studies. Across all categories coded, the average Cohen's Kappa was Cohen's K = 0.82. Table E1 in Appendix E presents a list of

computed Cohen's Kappa for each coded category. In instances where there were disagreements, discussions between the coders resulted in a common agreement being met.

Description of primary studies

Report and geographic characteristics

The final 32 studies included in the meta-analysis ranged in publication years from 1998 to 2018, with 2004 being the median year. A variety of publication types were selected, including journal articles, conference proceedings, a report, and one dissertation. The majority of the studies were journal publications (84%) from a publication source in either a technology or social science discipline (62%). Only five (16%) studies report being funded. Of those studies reporting the institution's geographic information (k =18), most were located in the Western region (50%) of the U.S. Table5 presents the frequency distributions for report characteristics.

Table 5

Frequencies of Report and Geographic Characteristics

Characteristic	*k*	*%*
Report Characteristics		
Publication Type		
Conference Proceeding	3	9
Dissertation	1	3
Journal	27	84
Report	1	3
Publication Source		
Social Science	9	28
Statistics Education	4	13

Table 5 (continued).

Characteristic	*k*	%
Technology	11	34
Other	8	25
Funded Research		
Yes	5	16
No	27	84
Location		
East	1	6
International	3	17
North	2	11
South	3	17
West	9	50

Sample/student characteristics

The total sample size across all 32 studies was comprised of 10,113 subjects (students). The majority of studies had undergraduate student samples (84%). Of the studies reporting information about students' gender (k =19), most samples had a majority of females (68%) and most studies had students who were mostly from the one gender (58%, k =24). Of those reporting disciplinary area (k =25), most courses were taught in social science (e.g. education, psychology, sociology, etc.) (56%), followed by natural science (e.g. physics, health) (16%), and applied sciences or humanities (business, criminal justice) (16%) disciplines. Table 6 presents the frequency distributions for student characteristics.

Table 6

Frequencies of Student Characteristics

Characteristic	*k*	*%*
Analyze		
Academic Level		
Undergraduates	28	85
Graduate students	4	12
Learners' Disciplinary Background		
Interdisciplinary	10	42
Same disciplines	14	58
Course Disciplinary Area		
Applied Sciences or Humanities (e.g. business, criminal justice)	4	16
Social Sciences (e.g. education, psychology, sociology, etc.)	14	56
Formal Sciences (e.g. math/statistics)	1	4
Natural Sciences (e.g. physical, health)	4	16
Multiple	2	8
Student Gender Composition		
Majority Female	13	68
Majority Male	4	21
Approximately Equal Number of Males & Females	2	11

Instructional design characteristics

Most studies described the course name as Introductory/Elementary Statistics (53%). The majority of classes were taught face-to-face (FTF) using either a lecture and lab or a lecture-only instructional delivery format (56%). The content area most often taught across studies was basic statistical concepts (ex: descriptive statistics, probability, sampling) (62%), followed by data analysis/ statistical tests (22%). As it relates to the learning goal of instruction, for most studies, it was learning statistical skills/concepts (59%), whereas for 22%, it was statistical literacy, thinking, or reasoning. Furthermore,

34% of studies used technologies with multiple learning goal functions. Those with single learning goal functions of technology were associated with automation of calculations (16%), collaboration and student involvement (16%), or simulation (16%). A variety of types of technologies were used (ex: statistical packages, digital games, tutorials, learning management systems (LMS), multimedia software, etc.). The technology types most frequently used were commercial statistical packages (22%), and LMS/CMS/web-based courses (22%), and stand-alone or web-based simulation/applet/visualization tools (13%). Technology was most often used for a semester or longer (66%). Furthermore, most technologies used were developed by the teacher/researcher (53%).

As it relates to the cognitive outcome function of technology, most were used for knowledge integration (53%), followed by knowledge organization (22%). In most cases, learners engaged with the technology individually (67%) and actively (directly) interacted with learning materials (75%). The type of feedback provided when interacting with technology was mostly immediate (50%). Of those studies providing information about the specificity of feedback (k =18), most technologies provided specific feedback (72%). Concerning the type of formative assessment employed while using technologies (k =30), most studies used either homework assignment/practice questions/activities (47%), followed by multiple measures (31%). Summative assessment measures consisted mostly of a teacher-made exam/test/quiz (72%), followed by multiple measures (22%), and standardized achievement/cognitive tests (6%). Additionally, only four studies (13%) used what was considered as authentic assessment summative evaluation approaches (e.g., assignment/project grade/presentation/demonstration/etc.), with the majority using

non-authentic assessments (72%). Table 7 presents the frequency descriptive information for instructional design characteristics.

Table 7

Frequencies of Instructional Design Characteristics

Characteristic	*k*	%
Course Name	3	9
Business statistics	1	3
Criminal justice research methods	1	3
Introduction to probability	2	6
Introductory social-science/social statistics	17	53
Introductory/elementary statistics	3	9
Medical/health science statistics	2	6
Psychology statistics	3	9
Research methods/research methods and statistics	3	9
Learning Goal		
Develop statistical literacy, thinking or reasoning	7	22
Learn stat skills/concepts	19	59
Understand purpose (logic) or process of stat investigations	6	19
Learning Goal of Technology Use		
Automation of calculations	5	16
Collaboration and student involvement	5	16
Investigation of real-life problems	2	6
Simulation used as teaching tool	5	16
Visualization of concepts	4	12
Multiple	11	34
Content		
Descriptive statistics, hypothesis testing	3	9
Distributions, probability, centrality, randomness	6	19
Data analysis/inferential statistics/statistical tests	7	22
Multiple basic concepts (descriptive statistics, probability, sampling)	16	50

Table 7 (continued).

Characteristic	*k*	%
Treatment Duration		
A semester or longer	21	66
Less than one semester	11	34
Design		
Instructional Delivery Mode		
FTF/Lab only	2	6
FTF/Lecture only	9	28
FTF/Lecture/Lab	9	28
Flipped/Hybrid/Blended/Distance Education	7	22
Online (All instruction online)	5	16
Develop		
Technology Design		
Institution hosted	6	19
Propriety (commercial)	9	28
Instructor/researcher designed	17	53
Media/Technology Type		
Commercial stats package	7	22
Digital game	2	6
Drill & practice or web-based tutorial/computer assisted learning	3	9
LMS/CMS/Web-based course	7	22
Multimedia/presentation software	3	9
Screencast tutorial/vodcast	3	9
Stand-alone or web-based simulation/applet/visualization Tool	4	13
Web information resource	3	9
Cognitive Outcome Function of Technology		
Information presentation	5	16
Information seeking	3	9
Knowledge integration	17	53
Knowledge organization	7	22

Table 7 (continued).

Characteristics	*k*	*%*
Implementation		
Learner Task (C-P)		
Assignments/Problem Solving	6	19
Lab exercises	8	25
Multiple	18	56
Learner Engagement (T-P)		
Cooperative/collaborative/collective	7	21
Individual	21	67
Mixed (students work alone & in groups)	4	12
Learner Control (T-P)		
Active/doer (learner w/ materials)	24	75
Expository instruction/receiver (learner w/ teacher)	1	3
Interactive/contributor (learner w/ peers)	2	6
Multiple	5	16
Scaffolding (T-P)		
Scaffolding present	16	50
No scaffolding	16	50
Feedback Type (T-P)		
Immediate	16	50
Not immediate	4	13
Both (Immediate and Not Immediate)	2	6
None	10	31
Specificity of Feedback (T-P)		
Non-specific (provide correct or incorrect feedback only)	4	22
Specific (provides feedback w/ detailed & specific response to behavior)	13	72
Technology Function with Concept (T-C)		
Computing(data analysis/diagnostics/ bootstrap) or graphing(distribution/outliers/models/centrality/spread)	5	16
Course mgt(collaboration)	6	19

Table 7 (continued).

Characteristics	*k*	%
data exploration	3	9
simulation(probability/variability)	6	19
Multiple	12	38
Evaluation		
Formative Assessment Measure		
Homework Assignment/Practice questions/Activities	15	47
Multiple	10	31
Quizzes/Test	7	22
Summative Assessment Measure		
Another achievement test (e.g. teacher made exam/test/quiz/chapter test)	19	59
Multiple (combined measures)	7	22
Standardized achievement/cognitive test	4	13
Other	2	6
Summative Evaluation Type		
Authentic Assessment (e.g., assignment/project grade/presentation/demonstration/etc.)	4	13
Non-Authentic Assessment (e.g., course grade/final/mid-term test/grade/exam/achievement test)	23	72
Both	5	16

Design, replicability, fidelity, and quality

The primary studies used two types of designs, independent groups post-test (59%) and independent groups pre-test and post-test designs (41%). Most studies provided descriptions of their instructional design process that could roughly be replicated (75%). Examining the implementation fidelity, the majority of studies used equivalent sets of learning materials for both treatment and control groups (78%). Furthermore, as it relates to the quality of studies, overall, most studies had an “unclear

risk of bias" (63%).When examining the distribution of studies in risk of bias categories (low, unclear, high) and across validity attributes (Internal, External, Implementation, Construct, Statistical Conclusion) for three out of the five attributes, more than half of studies fell in the low or unclear bias category (internal validity (66%), external validity (53%), and implementation fidelity (88%)). Furthermore, 50% of studies were associated with low risk of bias for construct validity and most studies were associated with high risk of bias for statistical conclusion validity (59%). Concerning the extent of risk of bias within studies, two studies, Mclaren (2004) and Wilmoth and Wybraniec (1998) had "high" risk of ratings across all validity attributes resulting in a "high" summary of risk rating. Whereas, only one study, Wang (1999) had a "low" summary of risk rating with "unclear" and "low" ratings across validity attributes. Table8 presents the frequency distributions of methodological and study quality characteristics coded from the primary studies.

Table 8

Frequencies of Method and Study Quality Characteristics

Characteristics	*k*	%
Methodological Characteristics		
Material Equivalence		
Same set of materials for experimental & control groups	25	78
Slight diff sets of materials but overall cover same content	7	22
Description of ID Process		
Mentioned with enough detail to roughly replicate	24	75
Mentioned with limited detail	8	25
Research Design		
Independent groups post-test (IGPT)	19	59
Independent groups post-test pre-test (IGPTPT)	13	41
Study Quality		
Summary of Risk Bias		
High	10	31

Table 8 (continued).

Characteristics	*k*	%
Low	2	6
Unclear	20	63
Internal Validity		
High	11	34
Low	1	3
Unclear	20	63
External Validity		
High	15	47
Low	17	53
Implementation Validity		
High	4	13
Low	16	50
Unclear	12	38
Construct Validity		
High	16	50
Low	16	50
Statistical Conclusion Validity		
High	19	59
Low	2	6
Unclear	11	34

Note: *K* represents the number of studies. Count less than $k = 32$ represent missing data.

Statistics achievement results

After transforming and computing 55 effect sizes from 32 primary studies (as described in the *Methods* section), 32 effect size estimates and their corresponding standard deviations measuring the effect of using technology compared to not using technology on student achievement were combined to compute a weighted average effect size for each study. These were included in the meta-analysis along with their standard

deviation. The methods used to compute the effect sizes are presented Table F1 of Appendix F. Furthermore, the interpretations of effect sizes as small, medium, and large that follow, correspond with the recommendations by Cohen (1969). According to Cohen (1969), effect sizes of 0.2, 0.5, and 0.8 represent small, medium, and large magnitudes of the effect.

The obtained (unweighted) effect sizes from individual studies that ranged from Hedges g = 0.64 to 1.10 and were used to calculate an overall effect size. The average unweighted standardized mean difference across k =32 studies was Hedges' g = 0.26, with a median of Hedges' g = 0.26. The distribution of unweighted effect sizes estimates (Figure 2) followed a symmetrical distribution with pseudo z skewness = 0.17 and pseudo z kurtosis = -0.85.

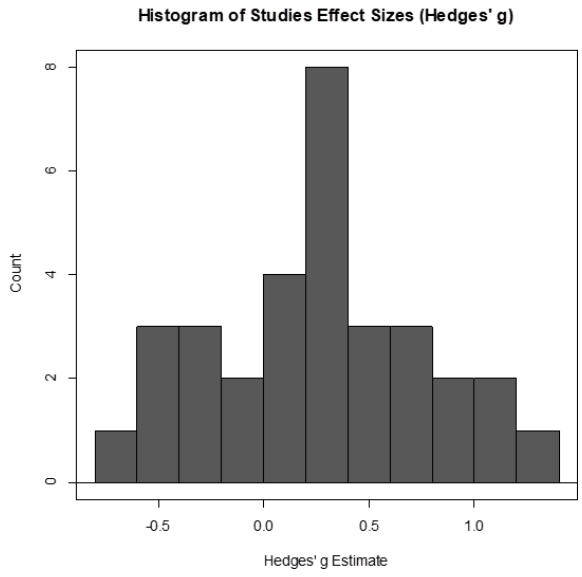

Figure 2. Distribution of Effect Sizes.

Histogram of 32 unweighted effect sizes based on statistics achievement outcomes showing a near normal distribution.

Outlier and influential diagnostics

An examination of outliers and influential cases for random-effects and mixed-effects models was conducted using criteria recommendations provided by Viechtbauer and Cheung's (2010). As shown in Figure 3, no outliers were found with standardized residuals within ± 1.96 standard deviations for the random-effects model. Furthermore, Cook's distance was used as a measure to examine influential cases. The results revealed that all Cook's distance values were within ± 3.13, therefore, providing an indication that there were no studies that would be considered influential. Additionally, a leave one-out analysis was conducted to examine if the observed significant effect would be non-significant when one study is removed, and the random-effect analysis conducted simultaneously. The findings did not reveal a significant impact on the overall effects size (based on a *Q*-statistic) when each study was removed one at a time and the random-effects model analysis was conducted on the remaining subset of studies. The results of the leave-one out analysis are reported in Appendix G, Figure G1.

Similarly, outlier and influence diagnostics were also conducted for the mixed-effects model which included the variables *Learning Task*, *Scaffolding*, and *Technology Function with Concept*. Standardized residuals and Cook's distances were examined. As shown in Table H1 in Appendix H, all measures were within the criteria recommendations provided by Viechtbauer and Cheung (2010). Standardized residual values were within ± 1.96, indicating no presence of cases that were outliers. Furthermore, Cook's distance values were within the ± 2.33 indicating no presence of overly influential cases.

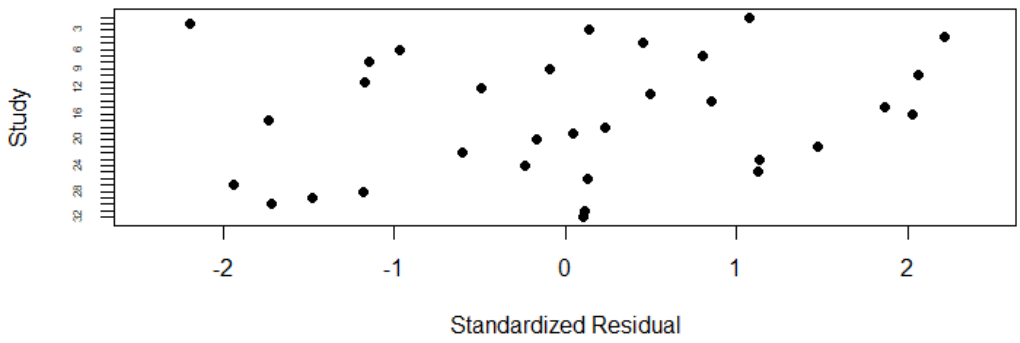

Figure 3. Plot of Standardized Residuals of k = 32 Studies.

A plot of the standardized residuals of effect sizes for individual studies showing residuals within |3| standard deviations (horizontal axis).

Research question one (part a)

The assumption was that the sample of studies are drawn from a larger population of studies having a distribution of true effect sizes that vary due to sources beyond sampling error alone (Cooper, 2017). Therefore, a random effects model was used to address the first part of research question one "*What is the overall average effect of using technology on statistics achievement?*" The weighted (inverse-variance) adjusted average standardized mean difference was Hedges g = 0.23, SE = 0.09, z = 2.63, and was statistically significant at p = .02, 95% CI [0.06, 0.41]. According to Cohen (1977), the estimated average effect of 0.23 corresponds to a small effect of technology use on statistical achievement. Therefore, on average, students who used technology had slightly but statistically higher statistics achievement by 0.23 standard deviations compared to students who did not use technology.

A forest plot provided further inspection and a visual representation of the distribution of weighted effect sizes around the overall average effect and their confidence intervals (see Figure 4). When visually examining the distribution of effect

sizes, the distribution was positively skewed, with more studies reporting effect sizes above the pooled effect size. Of the 32 standardized mean differences, 23 studies were in a positive direction and nine studies were in a negative direction. Also, the plot revealed that there were two studies, Hilton and Christensen (2002) and Lane and Aleksic (1998) with large sample sizes (by examining the size of the box shape) and thus, greater precision in their effect size estimate relative to other studies, also evidenced by the studies' small confidence intervals. Additionally, few studies had confidence intervals that exceeded the upper and lower limits of the confidence interval of the estimated pooled effect. This provided a need to further examine the heterogeneity of effect sizes to detect any potential outlier cases.

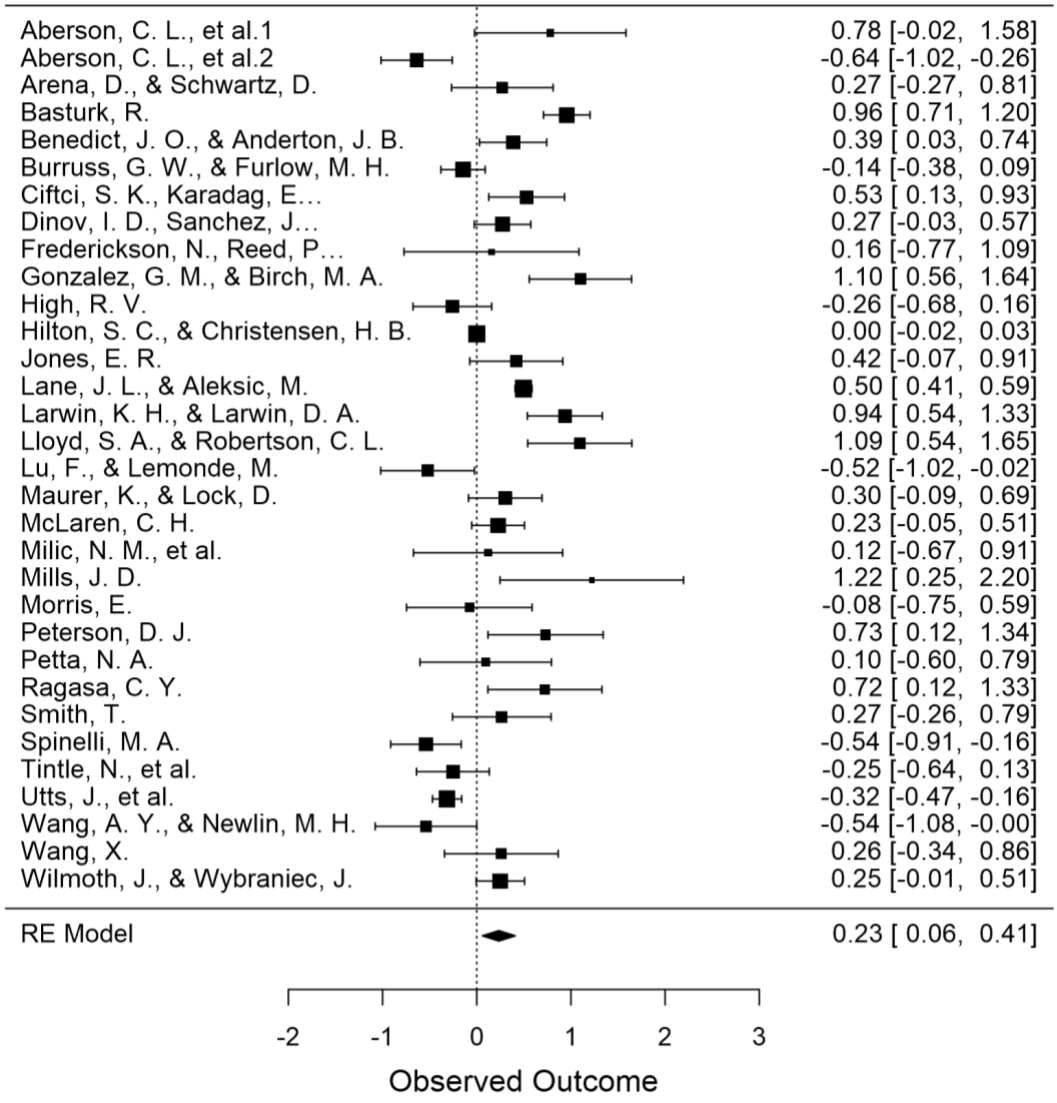

Figure 4. Forest Plot of Random-Effect Model for k = 32 studies.

A forest plot showing results of a random-effects model for 32 studies examining the effectiveness of technology use on statistics achievement. The figure shows the Hedges' *g* estimates in statistics achievement for individual studies using treatment (technology use) versus control (no technology use) conditions. Corresponding 95% confidence intervals are presented in brackets. The size of the squares represents each study's weighted contribution to the average weighted effect. The estimated weighted average effect is denoted by the diamond shape at the bottom of the figure.

Research question one (part b). A test of heterogeneity was conducted (random-effects model) to examine "*Is there a statistically significant difference in the variation of*

effect sizes around the estimated mean effect across studies?" The results revealed highly statistically significant heterogeneity, QE (31) = 306. 16, $p < .001$, indicating that the variation of effect sizes around the mean effect was greater than it would be by chance alone (Lipsey and Wilson, 2001). However, as Cochran Q statistic is sensitive to sample size (Lin, Chu, & Hodges, 2017), other heterogeneity measures were examined. These other measures provided further direction to examine heterogeneity. The amount of between-study variation was estimated at $\tau^2 = 0.20$, 95% CI [.11, .39] and the proportion of variation was found to be $I^2 = 93.56\%$. According to the suggestion by Higgins et al. (2003), an I^2 value of 75% or greater is an indication of considerable heterogeneity. Furthermore, as the results indicated heterogeneity, this provided further evidence supporting the need to conduct moderator analyses to examine if a portion of the heterogeneity could be explained by several study characteristics (potential moderators of the effect size).

Research question two

To account for some of the unexplained heterogeneity remaining, subgroup analyses were conducted to answer research question three "*To what extent do 24 study characteristics associated with phases of instructional design, moderate the effect of using technology on statistics achievement?*" Five variables (*Location, Student Gender Composition, Disciplinary Area, Learner Engagement, Specificity of Feedback*) were omitted from the moderator analyses due to missing data and therefore were not included in the reporting of the current findings.

The remaining 19 study characteristics can be discussed according to their association with phases of instructional design (ID) (ex: Analyze, Design, Develop, Implementation,

Evaluation). The subgroup analyses were conducted using a mixed-effects model with no-intercept; therefore, all levels of the factor (dummy coded) were included in the model. This provided results of the estimated mean effect for each factor and their respective confidence interval. Furthermore, the "Test of Moderators" was used to examine statistically significant differences between the pooled estimates of subgroups.

Design of instruction study characteristics. Separate mixed-effects analyses were conducted, and significant differences were found for moderators associated with each of the ID phases. The results and their associated statistics are reported for each factor examined in Table 9.

Analyze phase

At an alpha level of .05, the test of moderators was statistically significant for *Academic Level*, $Q_M(2) = 7.66$, $p = .02$; *Course*, ($Q_M(2) = 7.64$, $p = .02$; *Learning Goal*, $Q_M(3) = 11.74$, $p = .01$; and *Content*, $Q_M(4) = 9.49$, $p = .05$. This indicates that the estimated mean effect, jointly, for the levels of the factor, was not zero. Thus, at least one of the levels was a significant predictor of the effect size. On the other hand, the factor *Learning Goal of Technology* was not found to be a moderator of the size of effect with $Q_M(6) = 7.48$, $p = .28$. Subgroup differences for *Academic Level* yielded a small to medium mean effect favoring technology use for studies comprised of undergraduate students, Hedges' $g = 0.45$ ($p = .03$; $k = 28$; 95% CI [0.02, 0.39]). Therefore, undergraduate students using technology outperformed students not using technology on statistics achievement by 0.18 standard deviations, corresponding to a 95% confidence interval of possible higher true scores by 0.02 to 0.39 standard deviations favoring the treatment condition. Meanwhile, studies with graduate students was not significantly

associated with differences in the effect size, Hedges' $g = 0.18$ ($p = .10$; $k = 4$; 95% CI [-0.08, 0.97]).

Furthermore, differences in the estimated effect were found for studies where the introductory statistics course taught was not discipline specific (e.g. interdisciplinary focused). On average, students using technology in these studies had higher statistics achievement compared to those not using technology by 0.31 standard deviations, Hedges' $g = 0.31$ ($p = 0.01$, $k = 19$, 95% CI [.07, 0.55]). For *Learning Goal*, significant effects of technology use on statistical achievement were found within studies that used technology with the goal of learning statistical skills/concepts, Hedges' $g = 0.28$ ($p = .02$, $k = 19$, 95% CI [.05, 0.51]). Likewise, significant effects were found within the subgroup where students used technology to develop statistical literacy, thinking, or reasoning, Hedges' $g = 0.42$ ($p = .02$, $k = 6$, 95% CI [.07, 0.77]). For the variable *Content*, studies that covered content related to descriptive statistics and hypothesis testing during students' use of technology, on average, had highly significant medium effects on statistics achievement favoring technology use, Hedges' $g = 0.74$ ($p < .001$, $k = 3$, 95% CI [0.01, 1.38]). Those using technology had higher statistics achievement scores by 0.74 standard deviations compared to those not using technology. This corresponds to statistics achievement scores of students who did not use technology that are 69% to 79% below the mean achievement of students who used technology.

Design phase

As it relates to the "Design" phase, subgroup analysis revealed that the effect of technology use on statistics achievement was statistically significantly different for at least one level of the factor "*Treatment Duration"*, $Q_M(3) = 6.80$, $p = .03$. Small positive

effects were found for studies where technology was used for a semester or longer Hedges' g = 0.25 (p = .02, k = 20, 95% CI [0.04, 0.46]). The findings suggested that students using technology scored 0.25 standard deviations higher on student achievement than those that did not use technology when technology was used for a semester or longer. Meanwhile, no significant subgroup differences were found for the mode of instructional delivery, indicating that the method in which instruction was delivered was not associated with differences in the size of the effect of technology use compared to not using technology, $Q_M(5) = 8.99$, p = .11.

Develop phase

As it relates to the "Develop" phase, subgroup differences were found for *Technology Design* Q_M (3) = 7.72, p = .05. Studies in which the technology tool was developed by the instructor or researcher were, on average, associated with small to medium effects favoring technology use, Hedges' g = 0.30 (p = .02, k = 17, 95%, CI [0.06, 0.55]. This result indicated that when the technology was developed by a teacher or researcher, generally, students using technology had slightly higher statistics achievement by 0.30 standard deviations than those not using technology. This corresponds to a small effect on student achievement favoring technology use. Furthermore, results revealed that neither the type of technology used (e.g., statistician package/software, digital game, tutorial, learning management system, etc.) or the cognitive function of technology (e.g., information presentation or seeking, knowledge integration, knowledge organization) were not moderators of the effect size.

Implementation phase

Instructional design characteristics related to the "Implementation" phase that were found to have statistically significant subgroup differences included *Learner Task*, $Q_M(3) = 8.40$, $p = .04$; *Learner Engagement*, $Q_M(2) = 7.75$, $p = .05$, *Scaffolding* , $Q_M(2) = 9.26$, $p = .01$, and *Technology Function with Concept*, $Q_M(5) = 16.35$, $p = .01$. For *Learner Task*, studies where students completed multiple learning activities (ex: assignments, lab exercises, etc.) were associated with small to medium effects favoring technology use, Hedges' $g = 0.33$ ($p = .01$, $k = 18$, 95% CI [0.10, 0.56]). Students using technology, on average, scored 0.33 standard deviations higher on student achievement than their control group counterparts. Furthermore, a small to medium average effect (Hedges' $g = 0.36$) was found among studies where scaffolding was provided either by the student, teacher or technology tool, which was statistically significantly related to differences in the effect size ($p = .00$, $k = 16$, 95% CI [0.12, 0.61]). Whereas, studies where there was no scaffolding present had smaller effects that were not statistically significantly related to the size of the effect, Hedges' $g = 0.11$ ($p = .38$, $k = 16$, 95% CI [-0.13, 0.35]). As it relates to *Learning Engagement*, statistically significant effects on student achievement favoring technology use was found among studies where students engaged in cooperative, collaborative, or collective learning activities while using technology, Hedges' $g = .38$, ($p = .05$, $k = 7$, 95% CI [0.00 to 0.76]. Finally, for *Technology Function with Concept*, a significant positive effect on student achievement was found among studies where students used technology to cover concepts (e.g. probability/variability) through simulation (Hedges' $g = 0.42$), these were associated with statistically significantly positive effect on statistics achievement. Thus, on average,

students using technology had 0.42 standard deviation higher statistics achievement than students who did not use technology, ($p < .01$, $k = 6$, 95% CI [0.17, 0.68])

Evaluation phase

For characteristics associated with the "Evaluation" phase, subgroup analyses results revealed statistically significant effects were found for *Formative Assessment Measure* $Q_M(3) = 7.79$, $p = .05$ and *Summative Evaluation Type*, $Q_M(2) = 6.81$, $p = .03$. These results suggested that the type of formative assessment or summative evaluation were moderators of the average effect of technology use on statistics achievement. For *Formative Assessment Measure,* there was a significant effect on student achievement among studies that used multiple formative assessment measures favoring technology use, Hedges' $g = 0.34$, ($p = .03$, $k = 8$, 95% CI [0.03 to 0.65]). In these studies, students using technology had higher achievement scores by 0.34 *SD*, compared to those not using technology. Furthermore, for *Summative Evaluation Type,* findings suggested that among studies that used non-authentic assessments, there was a significant, yet small effect of using technology on statistics favoring technology use, Hedges' $g = 0.22$, ($p = .04$, $k = 23$, 95% CI [0.04 to 0.43]). Thus, students using technology had higher statistics achievement by 0.22 *SD*. This effect was smaller than for those using authentic assessment types, Hedges' $g = 0.28$, ($p = .10$, $k = 9$, 95% CI [-0.05 to 0.61]).

Table 9

Mixed-Effects Subgroup Analyses of Study Characteristics

Characteristic		*K*	*g*	*SE*	*z*	*p*	95% CI LB	95% CI UB	Heterogeneity Q_E	*df*	*p*	τ^2	*SE*	I^2
Analyze														
Academic Level	$Q_M(df = 2) = 7.66, p = .02$								263.38	30	<.01	0.19	0.06	93
	Undergraduates	28	0.21	0.10	2.21	.03	0.02	0.39						
	Graduate	4	0.45	0.27	1.66	.10	-0.08	0.97						
Course	$Q_M(df = 2) = 7.64$ $p = .02$								303.02	30	< .01	0.20	0.07	93
	Non-Discipline Specific Introductory Statistics	17	0.31	0.12	2.54	.01	0.07	0.55						
	Discipline Specific Introductory Statistics	15	0.14	0.13	1.10	.27	-0.11	0.40						
Learning Goal	$Q_M(df = 3) = 11.74, p = .01$								243.17	29	< .01	0.18	0.06	86
	Develop statistical literacy, thinking or reasoning	7	0.42	0.18	2.35	.02	0.07	0.77						
	Learn stat skills/concepts	19	0.28	0.12	2.42	.02	0.05	0.51						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*p*	95% CI LB	95% CI UB	Heterogeneity Q_E	Heterogeneity *df*	Heterogeneity *p*	Heterogeneity τ^2	Heterogeneity *SE*	Heterogeneity I_2
Learning Goal of Technology Use	Understand purpose (logic) or process of statistical investigations	6	-0.12	0.19	-0.60	.55	-0.49	0.26						
	$Q_M(df = 6) = 7.48, p = .28$								231.92	26	< .01	0.23	0.09	89
	Automation of calculations	5	0.20	0.23	0.86	.39	-0.26	0.66						
	Collaboration and student involvement	5	0.22	0.25	0.90	.37	-0.26	0.71						
	Investigation of real-life problems	2	0.49	0.40	1.23	.22	-0.29	1.26						
	Simulation used as teaching tool	5	0.42	0.24	1.75	.08	-0.05	0.90						
	Visualization of concepts	4	0.19	0.27	0.70	.48	-0.33	0.71						
	Multiple	11	0.15	0.16	0.93	.35	-0.17	0.47						
Content	$Q_M(df = 3) = 9.49, p = .05$								251.38	28	< .01	0.20	0.07	89
	Descriptive statistics, hypothesis testing	3	0.74	0.33	2.28	.02	0.10	1.38						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*p*	95% CI LB	95% CI UB	Heterogeneity Q_E	*df*	*p*	τ^2	*SE*	I^2
	Distributions, probability, centrality, randomness	6	0.25	0.21	1.22	.22	-0.15	0.66						
	Data analysis/inferential statistics/statistical tests	7	0.18	0.19	0.94	.35	-0.19	0.55						
	Multiple basic concepts (descriptive statistics, probability, sampling, randomness)	16	0.18	0.13	1.38	.17	-0.07	0.43						
Design														
Treatment Duration	$Q_M(df = 2) = 6.8, p = .03$								305.94	30	< .01	0.20	0.07	94
	A semester or longer	21	0.25	0.11	2.29	.02	0.04	0.46						
	Less than a semester	11	0.20	0.16	1.26	.21	-0.11	0.52						
Instructional Delivery Mode	$Q_M(df = 5) = 8.99, p = .11$								282.98	27	< .01	0.21	0.07	92
	FTF/Lab Only	2	0.27	0.36	.76	.44	-0.43	0.97						
	FTF/Lecture Only	9	0.32	0.18	1.81	.07	-0.03	0.66						
	FTF/Lecture/Lab	9	0.38	0.17	2.22	.03	0.04	0.71						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*p*	95% CI LB	95% CI UB	Heterogeneity Q_E	Heterogeneity *df*	Heterogeneity *p*	Heterogeneity τ^2	Heterogeneity *SE*	Heterogeneity I_2
	Flipped/Hybrid/Blended/Distance Education	7	0.03	0.20	0.13	.90	-0.37	0.42						
	Online (All instruction online)	5	0.10	0.23	0.43	.67	-0.35	0.54						
Develop														
Technology Design	$Q_M(df = 3) = 7.72, p = .05$								211.64	29	< .01	0.20	0.07	90
	Institution hosted	6	0.05	0.21	0.22	.83	-0.37	0.46						
	Propriety (commercial)	9	0.23	0.17	1.34	.18	-0.11	0.57						
	Instructor/Researcher designed	17	0.30	0.13	2.43	.02	0.06	0.55						
Media/Technology Type	$Q_M(df = 8) = 7.31, p = .50$								213.06	24	< .01	0.25	0.09	90
	Commercial statistical package/software	7	0.36	0.21	1.71	.09	-0.05	0.78						
	Digital game	2	0.27	0.40	0.67	.50	-0.52	1.05						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*P*	95% CI LB	95% CI UB	Heterogeneity Q_E	Heterogeneity *df*	Heterogeneity *p*	Heterogeneity τ^2	Heterogeneity *SE*	Heterogeneity I_2
	Drill & practice or web-based tutorial/computer assisted learning	3	-0.07	0.34	-0.20	.84	-0.73	0.60						
	LMS/CMS/web-based course	7	0.21	0.21	1.00	.32	-0.20	0.61						
	Multimedia/presentation software	3	0.29	0.31	0.94	.35	-0.31	0.89						
	Screencast tutorial/vodcast	3	0.11	0.32	0.35	.72	-0.51	0.73						
	Stand-alone or web-based simulation/applet visualization tool	4	0.31	0.27	1.18	.24	-0.21	0.83						
	Web information resource	3	0.25	0.35	0.72	.47	-0.43	0.93						
Cognitive Outcome Function of Technology	$Q_M(df = 4) = 6.38$, $p = .17$								240.15	28	< .01	0.22	0.07	89
	Information presentation	5	0.28	0.23	1.19	.23	-0.18	0.73						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*P*	95% CI LB	95% CI UB	Heterogeneity Q_E	Heterogeneity *df*	Heterogeneity *p*	Heterogeneity τ^2	Heterogeneity *SE*	Heterogeneity I_2
	Information seeking	3	0.25	0.33	0.74	.46	-0.41	0.90						
	Knowledge integration	17	0.23	0.13	1.82	.07	-0.02	0.49						
	Knowledge organization	7	0.21	0.20	1.05	.29	-0.18	0.59						
Implementation														
Learner Task	$Q_M(df = 3) = 8.40, p = .04$								176.75	29	< .01	0.19	0.06	89
	Assignments/problem solving	6	0.11	0.21	0.51	.61	-0.30	0.51						
	Lab exercises	8	0.12	0.18	0.67	.50	-0.23	0.46						
	Multiple tasks	18	0.33	0.12	2.77	.01	0.10	0.56						
Learner Engagement	$Q_M(df = 3) = 7.75, p = .05$								188.61	29	< .01	0.20	0.07	91
	Cooperative/collaborative/ collective	7	0.38	0.19	1.98	.05	0.00	0.76						
	Individual	21	0.21	0.11	1.94	.05	0.00	0.43						
	Mixed (students work alone & in groups)	4	0.07	0.27	0.25	.80	-0.47	0.61						

Table 9 (continued).

							95% CI		Heterogeneity					
Characteristic		*k*	*g*	*SE*	*z*	*P*	LB	UB	Q_E	*df*	*p*	τ^2	*SE*	I_2
Learner Control (T-P)	$Q_M(df = 4) = 9.11, p = .06$								282.60	28	< .01	0.2	0.07	94
	Active/doer (learner w/ materials)	24	0.17	0.10	1.64	.10	-0.03	0.37						
	Expository instruction/receiver (learner w/ teacher)	1	0.25	0.47	0.53	.59	-0.67	1.17						
	Interactive/contributor (learner w/ peers)	2	0.25	0.36	0.68	.50	-0.46	0.95						
	Multiple	5	0.58	0.24	2.38	.02	0.10	1.06						
Scaffolding	$Q_M (df = 2) = 9.26, p = .01$								263.91	30	< .01	0.19	0.06	92
	Scaffolding present	16	0.36	0.12	2.91	.00	0.12	0.61						
	No scaffolding	16	0.11	0.12	0.88	.38	-0.13	0.35						
Feedback Type (T-P)	$Q_M(df = 4) = 7.75, p = .10$								250.63	28	< .01	0.21	0.07	90
	Immediate	16	0.24	0.14	1.74	.08	-0.03	0.52						
	Not immediate	4	0.06	0.30	0.21	.83	-0.53	0.66						
	Both (immediate and not immediate)	2	0.63	0.40	1.55	.12	-0.16	1.42						

Table 9 (continued).

							95% CI		Heterogeneity					
Characteristic		*k*	*g*	*SE*	*z*	*P*	LB	UB	Q_E	*df*	*p*	τ^2	*SE*	I_2
Technology Function with Concept	None	10	0.22	0.14	1.51	.13	-0.07	0.50						
	$Q_M(df = 5) = 16.35, p = .01$								266.60	27	< .01	0.16	0.06	91
	Computing (data analysis/diagnostics/ bootstrap) or graphing(distribution/outliers/models/centrality/spread)	5	-0.19	0.20	-0.95	.34	-0.59	0.21						
	Course management (collaboration)	6	0.26	0.20	1.29	.20	-0.13	0.66						
	Data exploration	3	0.50	0.28	1.82	.07	-0.04	1.04						
	Simulation(probability/variability)	6	0.42	0.13	3.23	.00	0.17	0.68						
	Multiple	12	0.03	0.20	0.15	.88	-0.35	0.41						
Evaluation														
Formative Assessment Measure	$Q_M(df = 3) = 7.79, p = .05$								228.83	29	< .01	0.20	0.07	90
	Homework assignment/practice questions/activities	15	0.24	0.14	1.74	.08	-0.03	0.50						
	Quizzes/test	7	0.07	0.20	0.37	.71	-0.31	0.46						

Table 9 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*P*	95% CI LB	95% CI UB	Heterogeneity Q_E	*df*	*p*	τ^2	*SE*	I_2
Summative Assessment Measure	Multiple	8	0.34	0.16	2.15	.03	0.03	0.65						
	$Q_M(df = 4) = 6.93, p = .14$								199.90	28	< .01	0.22	0.07	91
	Unstandardized achievement test (e.g. Teacher made exam/test/quiz/chapter test)	19	0.21	0.12	1.75	.08	-0.03	0.45						
	Multiple (combined measures)	7	0.33	0.20	1.61	.11	-0.07	0.73						
	Standardized achievement/cognitive test	4	0.30	0.26	1.12	.26	-0.22	0.82						
	Other	2	0.05	0.35	0.14	.89	-0.64	0.74						
Summative Evaluation Type	$Q_M(df = 3) = 6.81, p = .03$								210.47	30	< .01	0.20	0.07	91
	Authentic assessment (e.g., Assignment/project grade/presentation/demonstration/etc.)	9	0.28	0.17	1.65	.10	-0.05	0.61						
	Non-authentic assessment (e.g., Course grade/ exam/quiz, etc.)	23	0.22	0.11	2.02	.04	0.00	0.43						

Note: QM =Test of Moderator, LB = Lower Bound, UB = Upper Bound.

Research question three

A mixed-effects model was used to conduct a multiple-variables meta-regression analysis to answer research question two, "*To what extent are implementation phase elements associated with interrelations between technology, pedagogy, and content predictors of the effect of using technology on statistics achievement?"* The factors (dummy coded) included in the model were those associated with the implementation phase of ID and representing inter-relations between technology, pedagogy, and content: *Scaffolding* (technology with pedagogy), *Learning Task* (pedagogy with content), and *Technology Function with Concept*. Overall, the omnibus test ("Test of Moderators") was statistically significant $Q_M(7) = 17.47$, $p = .03$. Therefore, this suggested that the standardized mean differences for all three instructional elements related to the interrelations between technology, pedagogy, and content were jointly different from zero. Furthermore, this indicated that at least one of the levels of the factors was significantly related to the effect size.

The "Test of Heterogeneity" was highly statistically significant, $Q_E(24) = 122.54$, $p < .01$, indicating the presence of heterogeneity with $\tau^2 = 17$, 95% CI [0.08, 0.41], signifying a slight reduction in the amount of between study variance from the reduced model by 0. 03. The proportion of variability not due to sampling error also reduced from $i^2 = 93.17\%$ to $i^2 = 85.59\%$. The remaining heterogeneity provided an indication that the variations in effect sizes could potentially be accounted for by other moderators.

The statistically significant result of the omnibus test suggested that at least one level of a factor (predictor) in the model was significantly related to the size of the effect. Given this, it was found that there was a significant effect for a level of the factor,

Technology Function with Concept. The results were similar to those found in the moderator analysis for this factor alone. Moreover, studies where the technology provided multiple functions (e.g. data exploration, simulation, graphing, etc.) for covering concepts, on average, reported significantly higher effects on statistics achievement, favoring technology use, $\beta = 0.69$, $p < .02$, 95% CI [0.90, 1.30]. Likewise, at an alpha level of .10, technologies used to cover concepts through data exploration were associated with higher effects on student achievement when students used technology compared to not using technology, Hedges' $g = 0.68$, $p < .06$, 95% CI [-0.04, 1.40]. The results of the mixed-effects multiple-variable meta-regression analysis is presented in Table 10.

Table 10

Results of Mixed-Effects Multiple-Variable Meta-regression Analysis for Technology Use on Student Statistical Achievement

					95% CI		Heterogeneity					
Model	Estimate	*SE*	*Z*	*p*	LB	UB	Q_E	*df*	*p*	τ^2	*SE*	I_2
Mixed-Effects $Q_M(df = 8) = 17.47, p = .02)$							122.54	24	< .01	0.17	0.06	83
Learning Task – Assignments/Problem Solving	-0.19	0.23	-0.81	.42	-0.65	0.27						
Learning Task – Lab exercises	-0.47	0.32	-1.46	.15	-1.11	0.16						
Learning Task – Multiple	-0.24	0.27	-0.90	.37	-0.76	0.28						
Scaffolding – Scaffolding Present	0.07	0.18	0.39	.70	-0.29	0.43						
Tech Function w/ Content – course management	0.49	0.34	1.41	.16	-0.19	1.16						
Tech Function w/ Content – data exploration	0.68	0.37	1.85	.06	-0.04	1.41						
Tech Function w/ Content – multiple	0.69	0.31	2.24	.02	0.09	1.30						
Tech Function w/ Content – simulation	0.38	0.33	1.15	.25	-0.27	1.03						

Note: QM = “Test of Moderator, QE = “Test of Heterogeneity”, LB = Lower Bound, and UB = Upper Bound” confidence intervals.

Research question four

Mixed-effects meta-regressions were conducted to examine "*To what extent do report or methodological characteristics of primary studies moderate the effect of technology use on statistics achievement?*" Report characteristics examined included: *Funding Status, Publication Year*; and methodological characteristics included: *Description of Instructional Design Process* and *Research Design*. An examination of the results for report characteristics revealed that there were no statistically significant subgroup differences for each of the moderators. Therefore, neither *Publication Status* nor *Funding Status* were significant predictors of the effect size.

As it relates to the methodological characteristics, a statistically significant difference was found between subgroups associated with the *Description of Instructional Design Process*, $Q_M(2) = 16.39$, $p < .001$. Studies that described their instructional design process with enough detail to roughly replicate it were associated with statistically significantly small to medium effects of technology use on statistics achievement, $g = 0.36$, ($p < .001$, $k = 23$). Thus, among replicable studies, on average, students using technology had higher statistics achievement by 0.36 standard deviations compared to those that did not use technology. Furthermore, no subgroup differences were found for *Research Design* $Q_M(2) = 0.52$, $p = .47$, indicating that whether the independent groups studies had post-test only or pre-test post-test designs was not statistically significantly related to the effect size. The statistical results of the report and methodological characteristics subgroup analyses are presented in Table 11.

Table 11

Mixed-Effects Subgroup Analyses of Study Characteristics

							95% CI		Heterogeneity					
Characteristic		*k*	*g*	*SE*	*z*	*p*	LB	UB	Q_E	*df*	*p*	τ^2	*SE*	I^2
Report														
Publication Status (Ref: Published)	$Q_M(df = 1) = .06, p = .81$								301.42	30	< .01	0.20	0.07	92
	Intercept	27	0.23	0.10	2.36	.02	0.04	0.42						
	Unpublished/Grey Literature	5	0.07	0.29	0.24	.81	-0.50	0.64						
Funded Research (Ref: No)	$Q_M(df = 1) = .10, p = .75$								299.54	30	< .01	0.21	0.07	93
	Intercept	5	0.25	0.10	2.49	.01	0.05	0.44						
	Yes	27	-0.08	0.24	-0.31	.75	-0.55	0.40						
Publication Year	$Q_M(df = 1) = .75, p = .38$													
	Intercept		-23.79	27.66	-0.86	.39	-78.00	30.42	256.19	30	< .01	0.20	0.07	91
	Year	32	0.01	0.01	0.87	.39	0.02	0.04						
Methodological														
Material Equivalence (Ref: Different sets of materials)	$Q_M(df = 2) = 05, p = .82$								256.19	30	< .01	0.20	0.07	91
	Intercept	9	0.25	0.10	2.36	.02	0.04	0.45						
	Slight difference but overall groups cover same content	23	-0.05	0.21	-0.23	.82	-0.46	0.37						

Table 11 (continued).

Characteristic		*k*	*g*	*SE*	*z*	*P*	95% CI LB	95% CI UB	Heterogeneity Q_E	*df*	*p*	τ^2	*SE*	I_2
Methodological														
Material Equivalence (Ref: Different sets of materials for T & C groups)	$Q_M(df = 2) = .05, p = .82$								256.19	30	< .01	0.20	0.07	91
	Intercept	9	0.25	0.10	2.36	.02	0.04	0.45						
	Slight diff but overall groups cover same content	23	-0.05	0.21	-0.23	.82	-0.46	0.37						
Description of ID Process (Ref: Mentioned with enough detail to roughly replicate)	$Q_M(df = 2) = 6.96, p < .01$								206.07	30	< .01	0.15	0.05	85
	Intercept	24	0.35	0.09	3.79	< .01	0.17	0.54						
	Mentioned with Limited Detail	8	-0.49	0.19	-2.64	.01	-0.86	-0.13						
Research Design (Ref: IGP)	$Q_M(df = 2) = .52, p = .47$								204.10	30	< .01	0.20	0.07	91
	Intercept	19	0.18	0.12	1.57	.12	-0.05	0.41						
	IGPrP	13	0.13	0.18	0.72	.47	-0.23	0.49						

Note: Ref =Reference group, IGP = Independent Groups Post-test, IGPrP = Independent Groups Pre-Test Post-Test, QM = "Test of Moderator, QE = "Test of Heterogeneity", LB = Lower Bound, and UB = Upper Bound.

Research question five

Study quality was examined by assessing the extent of risk of bias (ROB) inherit in the studies included in the meta-analysis that could influence the robustness of findings (overestimating or underestimating results), as well as the conclusions made. Composite scores were derived from a rating scale developed by the researcher that evaluated the extent of risk of bias as it related to the presence of favorable or unfavorable methodological characteristics covering internal, external, construct, and implementation validity concerns. The rating scale assessed risk of bias for each validity attribute based on whether there was enough evidence to support statements related to the concerns of validity. Appendix C presents the validity attributes and decision statements. Therefore, higher scores indicated evidence of the study's adherence to concerns of validity (low risk of bias – higher level of quality), while lower scores indicated lack of evidence of the study's adherence to concerns of validity (high risk of bias – lower level of quality). Evidence that validity concerns were addressed appropriately were associated with "Low" extent of risk bias (2 points), whereas, evidence that validity concerns were addressed inappropriately were associated with "High" extent of risk of bias (0 points). When there was insufficient evidence to make a conclusion about whether validity concerns were addressed, these were associated with "Unclear" extent of risk of bias (1 point).

Risk of bias was examined for each study across rating categories, as well as overall, for each validity attribute. Additionally, a *Summary risk of bias* was computed to assess the overall extent of risk of bias across studies, segmented by risk of bias category

ratings. Given this, mixed-effects subgroup analyses were conducted to address the research question, "*To what extent is the quality of primary studies a moderator of the effect of using technology on statistics achievement?*" An intercept model was used to examine contrasts between ROB rating categories, particularly between "High ROB" versus "Low ROB" and "Unclear ROB". Thus, "High ROB" was used as the reference category for all analyses.

Risk of bias across studies. A subgroup analysis was conducted on the *Summary Risk of Bias* variable and the findings were not statistically significant, $Q_M(2) = 0.46$, $p = .79$, indicating that overall, there were no statistically significant differences in the size of the effect of technology use across studies with rating categories of low, unclear, and high risk of bias. Similarly, no statistically significant differences were found for each of the five validity attributes. A plot illustrating the proportion of studies for each validity attribute is presented in Figure 5. Given the non-significant findings when assessing differences between the levels of risk of bias categories (subgroups) for *Summary of Risk Bias* and the other validity attributes, this suggests that observed differences in the size of the mean effect were not influenced by the extent of risk of bias found in individual studies. Thus, the extent of risk of bias is not a significant predictor of the size of the effect of technology use on statistics achievement. Most notably, this was apparent when comparing studies associated with high risk of bias to those associated with low and unclear risk.

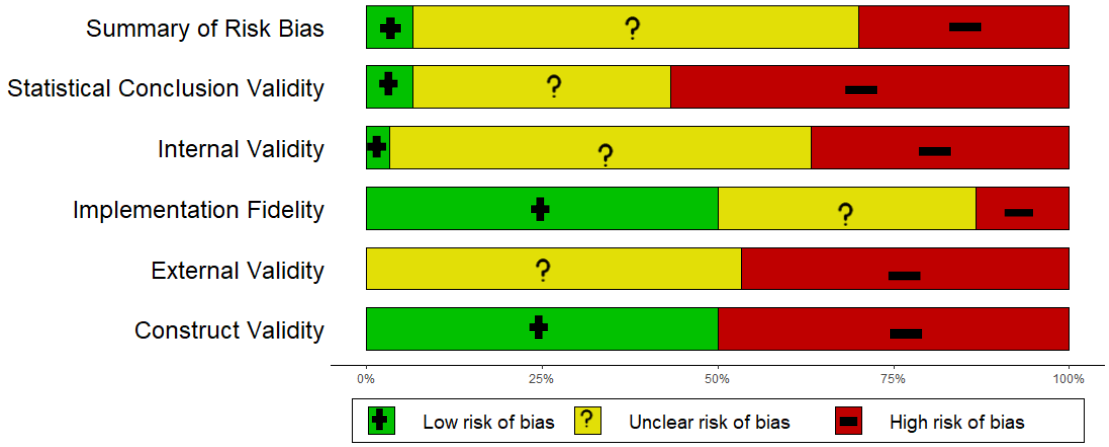

Figure 5. Extent of Risk Bias Across Studies.

Plot showing the extent of risk of bias across all studies (summary) in the meta-analysis by validity attributes. Risk of Bias category allocations are denoted by "+" (low ROB), "?" (unclear ROB), and "-" (high ROB). The plot was creating using the *robvis* tool by McGuinness (2019).

Risk of bias within studies. An evaluation of the extent of risk of bias inherent within studies revealed that out of the 32 studies represented in the meta-analysis, 10 studies were rated "high ROB" on three or more validity attributes; whereas, in contrast, the remaining 22 were rated "low" or "unclear" risk of bias on at least three or more validity attributes out of the five assessed. This is illustrated in Figure 6. Given this, subgroup analyses were conducted for the variable *Summary of Risk of Bias* to examine subgroup differences in the effect size between studies associated with "high ROB" and those with "low or unclear ROB". Mixed-effects subgroup analyses were conducted with the intercept included. There was no statistically significant difference found in the estimated mean size of effects between studies with high risk of bias and those with low or unclear risk of bias, $Q_M(2) = 1.91$, $p = .17$.

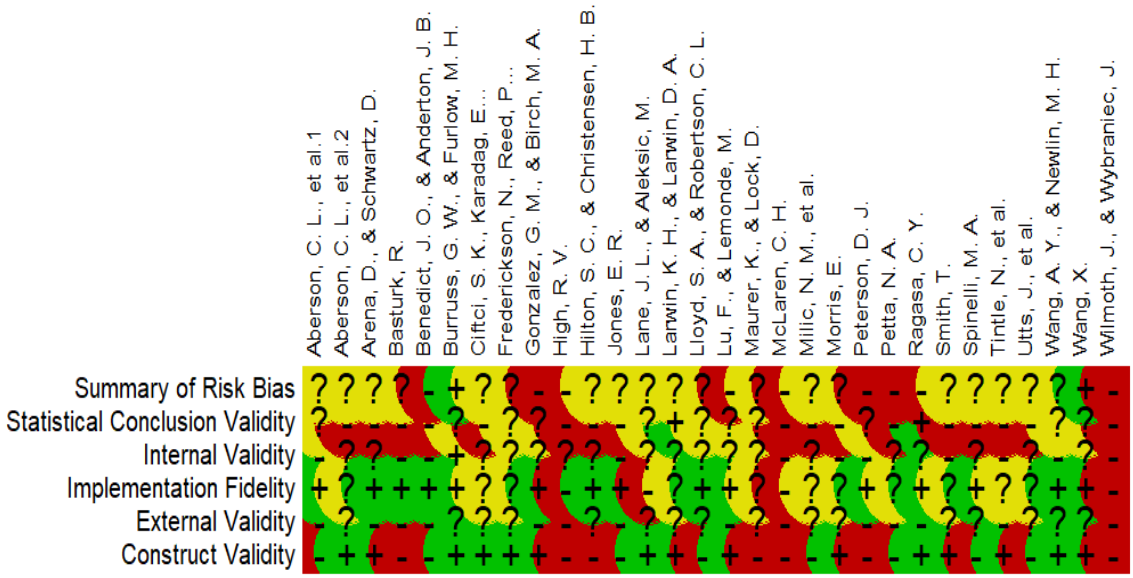

Figure 6. Risk of Bias Within Studies.

Plots showing the extent of risk of bias within each study segmented by validity attributes. Risk of Bias category allocations are denoted by "+" (low ROB), "?" (unclear ROB), and "-" (high ROB). The plot was creating using the *robvis* tool by McGuinness (2019).

Conclusions about extent of risk of bias

Considering the results obtained from the investigation of the quality of studies by examining the extent of risk of bias, it was found that risk of bias across categories, within and between studies, was not a statistically significant predictor of variations in the size of the effect. A qualitative interpretation of the results using recommendations from The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) (Atkins et al., 2004), would suggest an overall "unclear risk of bias" as most of the information about the extent of risk of bias across studies are from studies with "low" or "unclear" risk (67%). This suggests that meta-analysis conclusions should be

considered, bearing in mind that determinations about the extent of risk bias may raise potential doubts about meta-analysis results.

Publication bias

Publication bias is concerned with estimating the extent that missing studies might alter significant meta-analysis findings. This occurs when small studies are not published because of non-significant or negative findings, and therefore, are hidden from the literature. This potentially influences (overestimates) meta-analysis results (Sterne, et al., 2011). Indications of publication bias were examined in several ways. First, subgroup analyses were conducted to examine if there were significant differences in the effect sizes for published and unpublished/grey (e.g. doctoral dissertations/thesis, conference proceedings, reports). No statistically significant differences were found between the size of the effects of published and unpublished studies, $Q_M(1) = 0.06$, $p = .81$; therefore, this suggested that the size of effects was similar between published and unpublished studies. Additional examination of publication bias was conducted through the inspection of a funnel plot. The funnel plot provides a visual representation of the distribution of studies relative to their effect size (x-axis) and standard error (y-axis) about the pooled effect size. Therefore, it is expected that there is less dispersion across larger studies with small standard error (more precision) at the top half of the funnel; and more dispersion across smaller studies with large standard errors (less precision) at the bottom half. These results should reveal a symmetrical plot of scattered observations that resembles an inverted funnel. In contrast, when bias is present, there is a high concentration of studies on one side of the funnel compared to the opposite side Sterne, et al. (2011).

Visual inspection of the funnel plot resembled a nearly symmetrical plot. However, an examination of the corners of the funnel plot indicated the possibility that small studies with negative effect sizes could be missing. This was apparent in both plots as seen in Figure 7. Although the funnel plot provides a viable approach for estimating bias, it is subjective and can be difficult to interpret (Cooper, 2017). Therefore, Egger's regression test for plot asymmetry was conducted as a statistical approach for evaluating publication bias. Egger's test uses a linear regression method for testing publication bias, with the study's standard error (precision) as the independent variable and the effect size as the dependent variable (Egger, Smith, Schneider, Minder, 1997). Significant results ($p < .05$) indicate presence of publication bias. The result of Egger's test was non-significant, $p = .58$, therefore, suggesting that publication bias did not exist.

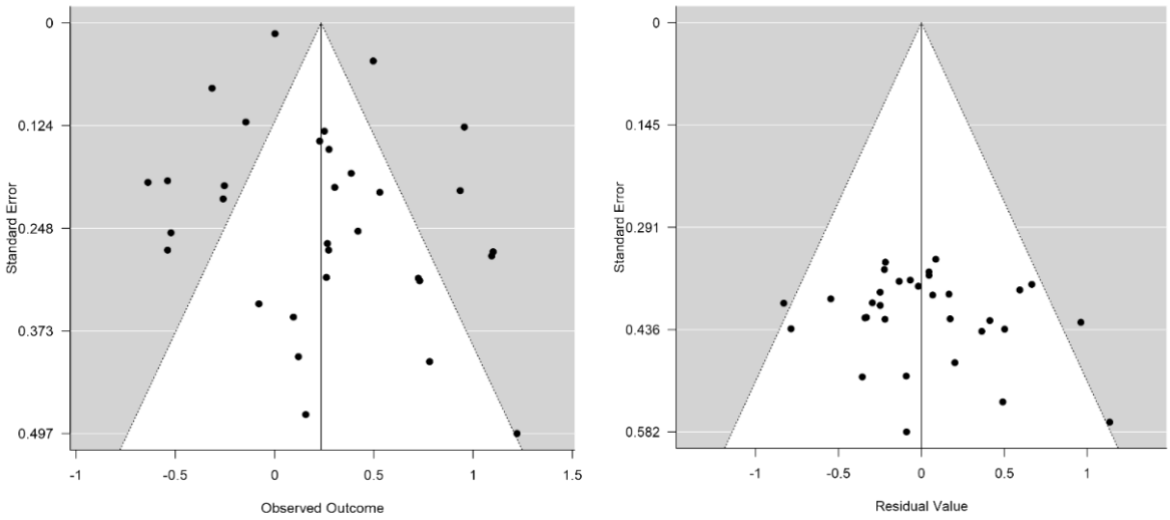

Figure 7. Funnel Plots of Individual Studies.

The funnel plot on the left represents the random-effects model (without moderator), while the mixed-effects model (with moderators) is plotted on the right. The funnel plots demonstrate observed effect sizes (left side) and residual values (right side) on the x-axis against their associated standard errors (y-axis) about the pooled effect size. A symmetric distribution of observation is an indication of no publication bias.

Conclusion

In this meta-analysis of 32 experimental or quasi-experimental studies related to technology use in statistics pedagogy, Overall, there was a small average effect of using technology compared to not using technology on statistics achievement favoring technology use (g = 0.23). An examination of moderating effects was conducted through mixed-effects subgroup analyses of 19 variables. This led to general findings that the learning goal of technology use, mode of instructional delivery, technology type, cognitive outcome function of technology, learner control, feedback type, and summative evaluation type had no appreciable relationship in explaining differences in the observed effect size. However, the remaining 12 factors were found to be significant moderators of the treatment effect (e.g. academic level, course, learning goal, content, treatment duration, technology design, learning task, learner engagement, scaffolding, technology function with concept, formative assessment measure, and summative evaluation type). On average, the subgroup findings favored the use of technology on student achievement by small to medium effects. Furthermore, an examination of report and methodological characteristics revealed no significant moderator effects for publication status, funding status, publication year, and research design. On the other hand, studies providing replicable descriptions of their instructional design were associated with significant moderator effects. Finally, visual and statistical results suggested no presence of publication bias; whereas, the extent of bias within and across studies was found to be mostly unclear.

CHAPTER V – DISCUSSION

The goal of the current study was to use a meta-analysis to investigate to what extent technology is effective in supporting introductory level statistics achievement and under what conditions it is most effective, considering elements related to the design of instruction. Tishkovskaya and Lancaster (2012) noted that for teaching to be effective, teaching and learning activities must be informed by pedagogical principles. Furthermore, it has been argued that current studies measuring the effectiveness of technology on learning are confounded by variables related to instructional design and teacher-related effects (Clark, 2001; Clark, 1994; Roblyer, 2005). Given these, the instructional design and TPACK/TPSK frameworks and constructivism learning theory were used to identify substantive study characteristics and examine their influence on students' learning of statistics in the technology-enhanced learning environment. Additionally, report and methodological study characteristics, as well as the extent of risk of bias were examined to assess the quality of studies in the meta-analysis and inform the relevance of meta-analysis conclusions.

Summary of findings

The current study used a meta-analysis approach to examine the effectiveness of using technology as an intervention in the introductory statistics classroom to support statistics achievement. Hedges' g was used to compute the effect sizes from 32 primary studies. Random-effects meta-analysis was used to answer research question one concerning the magnitude of the effect, as well as whether there was significant variation in the size of effects across studies. Results indicated a weighted average effect of 0.23

favoring the use of technology over not using technology on enhancing statistics achievement, corresponding to a small effect. Heterogeneity analysis indicated significant variation remaining in effect sizes across studies. Unexplained heterogeneity was examined through separate moderator analyses (analogues to ANOVA) to answer research question two regarding the extent that 24 study characteristics related to the design of instruction were moderators of the effect. A mixed-random-effects model was used, and findings revealed 12 study characteristics associated with each phase of the ADDIE instructional design framework that were significant moderators of the effect.

For the "analyze phase," these included students' *Academic Level*, *Course Type*, and *Learning Goal*. Significant positive effects favoring technology use was found for at least one level of these factors. These included average effect sizes of $g = 0.45$ for studies with undergraduate student samples, of 0.31 for discipline specific courses, and of 0.42 for studies with learning goals associated with statistical literacy, thinking, or reasoning and 0.28 for those with goals of learning statistical skills/concepts.

Furthermore, for the "design phase," *Instructional Delivery Mode* was not found to be a moderator of the effect. However, *Content* and *Treatment Duration* were significant moderators. Significant effects were observed among studies covering descriptive statistics or null hypothesis testing (0.74) and those using technology for a semester or longer (0.25). For the "develop phase," *Technology Type* was not a significant moderator, however, *Technology Design* was. Instructor designed tools were associated with significant effects on statistics achievement (0.30).

Next, for the "implementation phase" characteristics associated with *Learning Task, Learner Engagement, Scaffolding,* and *Technology Function with Concept* were significant moderators. Significant effects favoring technology use were associated with the use of multiple learning tasks (e.g. assignments, problem solving, lab exercises, etc.) (0.33), students' cooperative, collaborative, or collective engagement during learning activities (0.38), the use of scaffolding (whether by a tool or teacher) (0.36), and when technology was used to perform simulations when covering concepts (0.42).

Finally, "evaluation phase" characteristics associated with *Formative Assessment Measure* and *Summative Evaluation Type* were found to be moderators of the effect size. The use of a variety of formative assessment measures (e.g. assignments, quizzes, tests, etc.) was associated with a statistically significant effect (0.34). Meanwhile, studies in which authentic summative assessment was not used were significantly related to the size of effect on statistics achievement, favoring technology use (0.22). Although significant, the effect size was smaller than for studies using authentic summative assessment(which was not statistically significantly related to the size of the effect).

To answer the third research question, meta-regression analysis was conducted to examine the extent to which *Learning Task, Scaffolding,* and *Technology Function with Concept* jointly explained differences in the size of the effect of technology use on statistics achievement. The model was statistically significant, indicating that all three together, accounted for some of the variation in the size of effects; however, *Technology Function with Content* was found to be the only significant predictor. Among studies where the technology supported multiple functions for covering concepts (e.g. data

exploration, simulation, course management), there was a positive medium effect with higher statistic achievement among students using technology (0.68). Still, with heterogeneity remaining, the model did not account for all the observed variation in the size of effects.

To answer research question four, separate moderator analyses were conducted to examine if report and methodological characteristics were moderators of the effect. No significant results were found for report and methodological study characteristics associated with publication status, funded research, publication year, research design, and material equivalence. However, a significant effect was found for studies whose description of the instructional design process could be roughly replicated (0.36).

Next, to answer research question five, a mixed-effects analysis was used to examine if study quality was a moderator of the effect size. A researcher-developed scale for evaluating extent of risk of bias was used to assess study quality. Despite a non-significant finding for risk of bias within studies and across validity categories, meta-analysis findings were mostly based on studies with either unclear or low risk of bias (67%), with some having high risk of bias (31%). Therefore, it was concluded that overall, there was an unclear risk of bias associated with meta-analysis results. Finally, publication bias was examined both visually and statistically. Although visual inspection of the funnel plot suggested possible publication bias, sensitivity analysis for publication status and Egger's Test did not provide any indication of publication bias.

Literature synthesis of meta-analysis findings

In the current meta-analysis, a weighted average effect of technology use on statistical achievement of 0.23 was computed for 32 primary studies ranging within publication years, 1998 to 2018. This finding was similar to Schenker's (2007) meta-analysis findings (with which 17 studies overlapped with the current study). Schenker (2007) reported a statistically significant average effect of 0.24 when examining 46 studies published between 1985 – 2002. Meanwhile, other meta-analyses, (ex: Hsu (2003); Sosa, Berger, Saw, and Mary (2011), Larwin and Larwin (2011)) reported medium to large effects. The observed smaller average effect found in this study may be a result of the narrower inclusion criteria and smaller range of years that studies were published compared to previous meta-analyses. Given the former, only one meta-analysis (Hsu, 2003) restricted their inclusion criteria to introductory statistics courses, while the others included intermediate or advanced courses (Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011). Given the latter, the smaller range of publication years resulted in a smaller sample size from which findings were pooled. For example, Larwin and Larwin (2011) reported an effect size of 0.57 with studies covering a 50-year period; while, Berger, Saw, and Mary (2011) reported an effect size of 0.33 among studies covering a 31-year period. Additionally, the smaller observed effect might have been associated with the recent publication years. In their synthesis of scholarship on the use of information communication technologies in statistics education, van der Merwe and Wilkinson (2011) found that 64% of 162 articles were published between 2008 – 2009.

Additionally, considering concerns associated with technology integration in the classroom, it is plausible that the positive, yet, small mean effect size could be associated with educators' lack of knowledge, ability, or know-how in deciding the most-effective strategies for implementing technologies. Cobb (1992) and Pearl et al. (2012) addressed these, among others, as challenges faced by statistics educators when using technology to support student learning. Furthermore, Archer (2014) linked study quality concerns of implementation fidelity (of technology integration) to educators' levels of pedagogical, technological, and content knowledge. Another plausible explanation is that the magnitude of the effect size might be confounded by the presence of studies with high or unclear bias risk of bias (as observed in the current evaluation of study quality) which could lead to an underestimation of the true effect. The Cochran Collaboration (2011) pointed out that risk of bias inherit in studies can lead to underestimation or overestimation of meta-analysis results.

Furthermore, meta-analysis results indicated considerable heterogeneity, which led to moderator analyses to explain the remaining variation. This is reflective of the diversity of the field of research in which studies are conducted within a variety of instructional contexts and conditions (Garfield & Ben-Zvi, 2007). With this diversity and integration of technology in statistics classrooms, also comes challenges in teaching and learning statistical concepts (Pearl et al., 2012). Cobb and McClain (2004) emphasized the need for instructional design principles and learning theories to guide the implementation of activities for supporting students' statistical learning. Meanwhile,

Zieffler et al. (2008) suggested that post-secondary statistics educators can benefit from reviewing literature to gain a prescription for teaching and using technology.

Given these recommendations, the following discussion of findings from the moderator analyses is guided by the framework based on ID, TPACK, and constructivism presented in Table A-1 of Appendix A.

Analyze phase [context]: Assess learners, the context, and identify learning goals

Academic level. As it relates to learner characteristics, studies comprised of undergraduate students were associated with a significant and positive small to medium average effect of technology use on student achievement, favoring technology use. This is contrary to other meta-analysis findings (Larwin & Larwin, 2011; Schenker, 2007; Sosa et al., 2011) that reported positive effects among studies with graduate students. However, the current significant finding can be partially attributed to the current study's focus on introductory-level statistics courses which tend to have higher representation in undergraduate programs compared to graduate programs. This was represented in the current study as 88% (k = 28) of studies comprised of undergraduate student samples.

Course. The introductory statistics courses were examined by their disciplinary focus – non-discipline specific and discipline specific. Making up the majority (k = 17), non-discipline specific courses (e.g., courses with students from multiple disciplines and covering general statistics content) were associated with statistically significant effects on student achievement, favoring students using technology (0.31). Larwin and Larwin (2011) reported a similar finding for courses that offered general statistics to students from all disciplines (catch-all) (0.53). The observed positive and significant effect among

non-discipline specific courses could potentially be due to gains in achievement among students who come from various disciplines, with little or no prior experience with the subject. This is seen as reform through technology integration is grounded in a recognition of learning challenges faced by learners from diverse disciplines in introductory courses who tend to lack prior knowledge, experience anxiety, and have difficulty grasping concepts (Everson, Zieffler, & Garfield, 2008; Hassad, 2009). Chiesi and Primi (2010) commented that these challenges are even more prevalent among students with qualitative backgrounds. Therefore, technology integration supports diverse learners' ability to engage in their learning of statistical concepts (exploration, visualization, graphing, simulation, etc.), while addressing misconceptions that lead to the construction of new knowledge.

Learning goal. According to Chance et al. (2007), the selection of technology to support students' learning of statistical concepts should be based on a learning goal. This idea was supported by the current study's findings that revealed that learning goal was a significant moderator of the magnitude of the effect on statistical achievement. A significant and larger average effect was found when the learning goal was to develop statistical literacy, thinking, or reasoning, followed by the goal of learning statistical skills and concepts. These findings are relevant when considering that reform efforts largely emphasize the use of technology to achieve positive learning outcomes related to students' development of statistical literacy, thinking, or reasoning, as well as understanding basic or central statistical concepts (Cobb and McClain, 2004; Gaise, 2016).

Design phase [instructional delivery strategy]: What is it to be learned and how?

Content. Regarding the influence of characteristics associated with the design phase, when content covered was related to either descriptive statistics or null hypothesis testing, students using technology had significantly higher achievement scores compared to those not using technology (0.74). This finding can possibly be explained by the difficulty experienced by introductory learners in grasping foundational statistical concepts. This is compared to the other content areas that were not found to be significantly related to the effect size (ex: probability, data analysis, inferential statistics, etc.). In their study surveying 102 students enrolled in an undergraduate biostatistics course, Xu et al. (2014) found that students reported confidence intervals and hypothesis testing as the most difficult concepts to grasp. Tools such as tutorials support scaffolding of students' learning of these concepts (e.g. informal hypothesis testing) and provide immediate feedback that potentially helps students mitigate gaps in understanding (Aberson et al., 1997).

Treatment duration. Results revealed a statistically significant difference in the size of effect for studies where technology was used for longer than one semester. Students using technology had slightly higher achievement (0.25) than those not using technology. This was similar to Larwin and Larwin's (2011) and Sosa et al.'s (2011) findings of significant positive effect when using computer-assisted instruction for frequent and longer periods of time. It is possible that the observed significant effect is due to a time-to-task phenomena in which longer durations with repeated exposure and

practice being associated with students' engagement in learning activities, results in greater learning outcomes.

Develop phase [technology]: Produce or acquire instructional material

Technology type. Meta-analysis studies in the literature have found both statistically significant results (positive) (Hsu, 2003; Schenker, 2007; Sosa et al., 2011) and no difference (Cobb & McClain, 2004; Garfield et al., 2012) in the size of effect on student achievement when examining the influence of different types of technology. In the current study, the technology type (ex: statistical software, digital game, drill & practice, LMS, tutorials, etc.) was not found to be a statistically significant moderator of the effect size. This could be attributed to various reasons. According to GAISE, regardless of the type of tool used or its function for generating output, the basis for using technology should be in its usefulness to enhance students' conceptual understanding and learning. Furthermore, the non-significant finding can be viewed as a support of Clark's (1994) position that technology alone does not influence learning. This was evident as other features of the classroom environment, related to content and pedagogy, were found to be significant moderators of the effect size. Specifically, the interrelation between the function of technology and the concept covered was significant when the technology used had multiple functions (ex: computing, graphing, simulation, course management, etc.) when concepts were covered (e.g. probability, variability, centrality, etc.). Maker and Sousa (2014) described this as teachers' technological statistical knowledge, where statistical content is presented with the appropriate use of technological tools This finding is also associated with recommendations by Cobb and McClain (2001) for integrating

constructivist approaches to reform-based teaching that integrate technology whenever possible for automation of calculations and graphing. For example, in Basturk's (2005) study, students in the treatment group used real data sets to learn about measures of central tendency. This study was conducted in a lab using a statistical software (SPSS) for computations, to perform data analysis, generate outputs, and interpret data. Lastly, the inability to detect significant effects may be a result of the small sample size that comprised the meta-analysis.

Technology design. Studies in which the technology used was designed by the instructor were associated with significant effects on statistics achievement, favoring students using technology. The size of the average effect was larger for these studies than for those where technology was designed by the institution (0.05) or commercially (0.23). The significant effect could be associated with instructors' understanding of learning needs of students and therefore, being able to customize or select technologies or to implement features that would meet those needs. In contrast, institutional/commercial tools, applications, or software, are designed to meet learning needs of a general population of learners.

Implementation phase: Use of materials and strategies to deliver instruction

Learning task (content and pedagogy). The selection of instructional activities to support students' mastery of content covered was found to influence the size of the effect. Statistically significant positive findings were observed among studies in which students engaged in multiple learning tasks. The effect was larger than for studies where students completed one learning task (lab exercise or assignments/problem solving). These

findings are not surprising as the diversity of learning tasks provide different methods for students to learn, practice, and demonstrate mastery of content. This also provides differentiated learning experiences that appeal to different learning styles/characteristics/needs.

Learning engagement, scaffolding (technology and pedagogy). Studies in which student engagement involved cooperative, collaborative, or collective learning, were associated with significant effects on student achievement. The observed larger magnitude of effect for this subgroup (0.38) relative to individual learning (learner and technology) (0.21) and working individually and in groups (0.07) can potentially be explained by greater opportunities for learning to occur through collaboration with others. Additionally, the moderating effect on student achievement can be understood through social interdependence theory. According to this theory, when students engage in cooperative activities where there is an individual and collective stake to demonstrate the accomplishment of a task, they are more likely to exert greater effort in ensuring successful outcomes (Roseth, Garfield, & Ben-Zvii, 2008). Research investigations have reported positive effects on statistics achievement tests and scores when students engaged in cooperative learning (e.g. working in pairs and small groups) (Zieffler et al., 2008).

Furthermore, when scaffolding was embedded in learning (by the tool or instructor), this was associated with larger effects on statistics achievement favoring technology use (0.36) than when no scaffolding was present (0.11). The significant effect observed from the inclusion of scaffolding can be attributed to additional guidance provided by the tool that reinforces the progression of learning based on patterns in

student performance and understanding. Students using technology also benefit from scaffolding provided by the tool and by the instructor. Hassad (2011) explained that in constructivist learning environments, active learning and scaffolding activities lead to learners' exploration of concepts and construction of meaning applied to new knowledge.

Technology function with concept (technology and content). No known meta-analysis study to date has specifically examined the influence of instructional elements related to the synergy between technology and content on statistics achievement. When examined alone, it was found that the fusion between technology and content was significantly related to the size of effect on statistics achievement when simulation technology was used to learn concepts related to probability and variability. This finding is reasonable, as simulation tools have been found effective in enhancing students' understanding of abstract fundamental concepts such as distributions, probability, and chance as learners are able to visual and explore data that represent these concepts (Chance et al., 2007; Garfield & Ben-Zvi, 2007). In their study, Lane and Tang (2000) reported higher effects on learning associated with statistical reasoning when students used a simulation tool to learn about randomness and other statistical concepts compared to those who use a traditional textbook. Also, the use of simulation tools has been associated with learners' ability to perform well on problems related to probability (Garfield, 1995).

However, when the interrelations between learning task, scaffolding, and technology function with concept were examined together as predictors of the size of effect, a statistically significant result was found among studies where multiple functions

of technology were used to cover multiple concepts. The effectiveness of this approach was demonstrated in Garfield et al.'s (2012) study that used a curriculum called Change Agents for Teaching and Learning Statistics (CATALST) to examine its effect on developing statistical thinking among undergraduates in an introductory statistics course. The curriculum fused content, pedagogy, and technology with the use of Tinkerplots and Fathom software that enabled simulation and modeling approaches through which students investigated concepts of chance, randomization, and resampling. This resulted in students' increased ability to make statistical inference. Furthermore, reflecting on their findings, the authors concluded that **"**Students can be taught to "really cook" [not just follow recipes] by using a modeling and simulation approach to statistical inference along with TinkerPlots™ software" (Garfield et al., 2012, p. 896). Furthermore, the use of collaborative or collective engagement and scaffolding provide opportunities for both students and teachers to contribute together in knowledge construction (Garfield & Ben-zvi, 2008).

Evaluation [pedagogy]: Monitor and assess the effectiveness of instruction.

Moderating effects were associated with *Formative* and *Summative* assessment practices used to monitor student learning and the effectiveness of instruction. Formative assessment methods that involved the use of a variety of assessment measures (ex: homework assignments, practice questions, activities, tests, and quizzes) were associated with significant effects on statistics achievement favoring technology use. This finding aligns with recommendations from leading researchers in the field that urge assessment practices that include a variety of methods to evaluate student learning so that feedback

can be provided to enhance learning outcomes (GAISE College Report ASA Revision Committee, 2016; Garfield & Ben-zvi, 2007). Furthermore, the significant finding for summative evaluation measures that were not authentic could be reflective of the large proportion of studies that used this type of assessment measure (72%), compared to the fewer that used the recommended authentic type of assessments (e.g. minute paper, projects, performance task, etc.) (Garfield et al., 2008). On the other hand, the non-significant finding for authentic assessments could be reflective of evidence that suggests that authentic assessment approaches have a greater influence on affective outcomes, such as student's attitude toward statistics, than on cognitive outcomes (Hassad, 2014).

Report and methodological characteristics

The observed non-significant finding for publication status and funded research could have resulted from uneven distributions of observations in the groups compared. Most studies (84%) were published compared to those that were unpublished/grey literature. Similarly, the majority of studies (84%) were not funded. Furthermore, non-significant findings for material equivalence provides an indication of the fidelity of the implementation of treatment and control across studies, which consisted primarily of studies that used the same sets of materials in the treatment and control groups (72%), compared to those that had a slight difference but overall covered the same content.

Study quality

The extent of risk of bias was evaluated as a measure of study quality. Overall, findings revealed that the majority of studies in the meta-analysis had unclear risk of bias.

This finding is consistent with concerns that have been raised about the quality of research in the field.

Contributions and implications

The current study aimed to advance research by addressing the call for more evidence on the effectiveness of integrating technology in introductory statistics education as an instructional approach for supporting students' learning of statistical concepts (Chance, et al., 2007; Eichler & Zapata-Cardona, 2016; Hassad, 2014). This was accomplished, also recognizing that efforts to understand optimal ways to enhance student learning using technology is among the leading priorities for connecting research to practice (Pear et al., 2012). The current study went beyond describing the tools that are most effective in supporting learners' statistics achievement. Rather, considerations were made about the potential influence of factors associated with the learning context, content, and pedagogical strategies employed. This is consistent with the urge for a comprehensive examination of the learning contexts in which technology can be effective (Pearl et al., 2012; Roblyer, 2005). The findings of the study were framed using a framework that is a first meta-analytic attempt at potentially guiding statistics educators through the instructional design process when incorporating the use of technology to support student learning. This was based on an attempt to provide a prescription of the most effective strategies for integrating technologies in the introductory statistic classroom.

Overall, the current meta-analysis yielded a small average standardized mean effect of 0.23. Although small, the practical significance of the effect size can be gained

by recognizing that the standardized difference can be related to the area under the curve of 9.1%. In the context of students' statistics achievement, this could translate to a move of one letter grade over an academic period. Furthermore, concerning research priorities toward understanding the impact of technology on student assessment, Pearl et al. (2012) noted that research evidence would also help students, statistics educators, and administrators understand the cost implications of investing in technologies. This provides a reminder that the practical significance of an educational intervention depends on considerations of both the benefits and costs associated with implementation (in the current case technology).

The current study's findings revealed that technology type was not a moderator of the effect, which is consistent with Clark's (1994) claim that technology is merely a vehicle for transferring knowledge and alone does not influence learning. Still, findings revealed moderating impacts among 12 characteristics associated with each instructional design phase of the ADDIE framework. Each phase corresponded with unique instructional design objectives that can provide a guide for educators' implementation of findings. Additionally, with awareness of the need for change in how statistical concepts are taught and how students learn statistics (Garfield & Ben-Zvi, 2008), TPSK and constructivist learning ideologies were integrated into the conceptual framework used in the current study. To some extent, this helped to (indirectly) conceptualize and infer the technological pedagogical statistical knowledge (Lee & Hollebrands, 2008) possessed by researchers/instructors. . For example, moderating effects were found when activities were implemented while considering technology affordances with content and pedagogy

(*Learning Task, Learner Engagement, Scaffolding, Technology Function with Concept*). These findings provided basis for further implication about the extent and influence of using constructivist learning strategies on students' statistics achievement (Hassad, 2011; Zieffler et al., 2012).

Finally, prior criticisms concerning the lack of consideration of quality of research evidence among systematic reviews was addressed. This was accomplished through an investigation of the quality of primary studies based on the extent to which risk of bias was inherent within and across studies. Upon assessing the quality of the evidence among primary studies included in the current meta-analysis, the conclusion was that there is unclear risk of bias. This suggests that the practical reasonableness of current meta-analysis findings should be considered in light of potential bias that may underestimate or overestimate the true findings.

Limitations

Developing a model guided by instructional design principles provided a framework for understanding the contextual and pedagogical elements that interplay with the use of technology for teaching and learning statistics and that lead to desirable learning outcomes. However, the current study was limited in the operationalization and selection of instructional design variables. This was due to a lack of reporting and detailed descriptions of the research setting and contexts among primary studies. This resulted in the collapsing of some variables into the most meaningful categories possible.

Though robust, the meta-analysis approach has certain limitations. First, researchers are restricted by the availability of implementation and methodological data

or information provided in the study. Given this limitation, substantive learner characteristics data such as prior knowledge and level of self-direction could not be collected. Jung and Lee (2015) stressed the importance of considering learners' characteristics, preferences, and technology acceptance when assessing the effectiveness of technology use on learning. This would help to ensure that students are not hindered in their ability to meet learning goals. For example, Schmid et al (2014) found that the effects of technology use on post-secondary student achievement was higher for those in programs such as humanities, education, and language. Similarly, Vo, Zhu, and Diep (2017) point to factors associated with learners' characteristics and prior achievement as significant predictors of achievement in learning. Additionally, due to limitations of access to variables, the current study used the PTACK framework to make inferences about instructors' level of knowledge. However, the availability of observed instructors' PTACK data would have enhanced the analysis and interpretation of findings for more meaningful practical application.

Next, the process of article retrieval was comprehensive, which included consultation from a qualified librarian for the identification of relevant keywords, as well as the use of various combinations of key words and multiple database sources. However, there is a possibility that some relevant studies may not have been included in the meta-analysis, resulting in publication bias. This can result in the failure to utilize the entire domain of relevant keywords in the search process or the lack of reporting of non-significant findings associated with the file drawer problem.

The sample size from which meta-analyses were conducted was $k = 32$ studies. Furthermore, separate mixed-effects subgroup analyses were conducted to examine potentially moderating effects. When conducting subgroup analyses among levels of the factor, in some cases, this yielded smaller number of observations within one or more levels of the factor relative to other levels. Therefore, the low observation counts in the subgroups could have introduced some bias and reduced the robustness of the subgroup analyses.

Additionally, the current study used multiple measures for examining publication bias which provided a method of triangulation. This revealed a consistent pattern of no publication bias from statistical analyses which tends to strengthen confidence in conclusions. However, there were unequal numbers of observation in subgroup levels of *Publication Bias*. The number of unpublished/grey literature was relatively small compared to published studies. Therefore, a larger sample size would provide greater power that might result in detecting significant differences. Similarly, the conclusion made about the quality of studies in the current meta-analysis was that there was an unclear risk of bias. This was due to a relatively large proportion of studies that did not provide sufficient evidence to make a determination about the extent of risk of bias across and within studies. Moderator analysis suggested no differences across risk of bias categories. However, given a larger sample size, significant differences might exist between studies with low and high extent of risk of bias.

Finally, the meta-analysis approach provided a method for examining the effectiveness of technology use as an intervention for supporting statistics achievement in

the introductory classroom. This was conducted using a random sample of relevant studies from the literature. Although inclusion and exclusion criteria limited studies to those with experimental and quasi-experimental designs, findings of moderator analyses should be interpreted as correlational. As such, causal inferences can only be supported through direct manipulation of the study characteristics that were examined as potential moderators in the current study.

Recommendations and conclusions

Overall, through an instructional design lens, findings from the current study provide a foundation for understanding the potential impact of technology use in supporting students' learning of statistics in the introductory classroom. According to Pearl et al. (2012), the effective use of technology on statistics achievement is highly dependent on learners' interaction and engagement with the tool and others, as well as the scaffolding provided to guide the learning experience. This was evident in the current findings of both moderator analyses and meta-regression that examined the moderating effect of study characteristics that were related to the design of instruction. Findings supported the positive influence of instructional design characteristics associated with the inter-relationships between technology, pedagogy, and content on students' statistics achievement. These significant effects were also associated with constructivist learning practices that align with GAISE recommendations (GAISE College Report ASA Revision Committee, 2016) (e.g. cooperative/collaborative/collective learning, multiple learning tasks, formative assessment approaches, etc.). Additionally, findings were reported according to phases of instructional design. This provides educators and

practitioners a first overview of types of curriculum design decisions at each phase that might influence the effectiveness of students' learning experience when using technology.

Despite these findings, more insight is needed on the sequencing of instructional design elements that jointly influence students' achievement of reform-based student learning outcomes. Future meta-analysis research should seek to expand on the use of model-based frameworks that examine and test multivariable relationships among elements of instructional design in technology-enhanced introductory statistics classrooms. Specifically, research could test the applicability of the study characteristics that were found to be significant moderators in the current study. This would enable empirical investigations that examine how the sequence of instructional design activities (related to technology, pedagogy, and content) predict statistics learning outcomes (e.g. statistical literacy, thinking, and reasoning). In turn, this would provide meta-analysis researchers measurable constructs and variables to examine plausible instructional design models that lead to effective instruction in statistics education. These types of studies could contribute valuable insight about the associations between technology use and the achievement of learning outcomes in introductory-level statistics education.

However, this would require that primary studies incorporate these elements in the instructional design process with a clear description in their reporting of findings and directly examine their association with study outcomes. Primary researchers would need to operationalize, and measure constructs related to the elements or components that align with phases of the design of instruction. For example, researchers may use instructional

design models such as the ADDIE, Dick and Carrie, or other ID frameworks to inform decisions about how to design the instructional contexts in which the assessment of learning will occur. Additionally, it would require an interdisciplinary approach to this research that integrates recommendations from statistics education and educational technology literature to identify appropriate theories and frameworks.

In the current study, several characteristics (e.g. location, learners' disciplinary background, student gender composition, specificity of feedback, etc.) associated with the *Analyze* and *Implementation* phases were not included in the subgroup analyses because only a few studies provided information about these characteristics. This information would especially help in further understanding different learner profiles in the introductory statistics course that may be associated with particular instructional design strategies that ultimately lead to effective outcomes. The availability of this information would help support technology implementation decisions, enable statistical investigations, as well as provide relevant data that can be used in future meta-analysis research. Meta-analysis researchers would have available measures and variables to develop and test viable technology-infused instructional design models that lead to effective instruction in statistics education. These types of studies could contribute valuable insight to practitioners and could potentially be helpful in developing statistics educators' technological, pedagogical, statistical, knowledge. Additionally, findings from these studies would be useful for informing and staying abreast of best practices of technology use in statistics education.

Finally, Garfield and Ben-zvi (2007) raised concerns about the lack of high-quality measures used to assess student learning outcomes among quantitative studies. In a similar matter, leading researchers have criticized the common use of final exam or course grades, which pose challenges in substantiating the reliability and validity of findings (Garfield and Ben-zvi, 2007). This concern was reflected in the current study as the majority (80%) of studies used some form of final grades or quiz and exam scores as a summative assessment measure. There continues to be a dire need for evaluating the validity of findings reported in empirical research that examines the effectiveness of technology use in statistics education (Garfield & Ben-zvi, 2007; Zieffler et al., 2008). Furthermore, there is a need for studies with strong research designs and that use or report the psychometric properties of measurement instruments (Zieffler et al., 2008). The current study provided a first attempt at addressing the gap in examining the quality of empirical evidence in the literature by looking at the extent of risk of bias. Overall, it was concluded that the extent of risk of bias was unclear and limitations concerning the validity of the instrument were acknowledged Therefore, this should be taken into consideration when interpreting the current findings. Future meta-analysis research should focus on further developing valid instruments or evaluation rubrics that can be used to assess study quality among studies examining statistics learning outcomes in technology-enhanced classroom. These should also take into consideration discipline-specific threats to validity as noted by Clarke (2001) (e.g. fidelity of technology implementation, equivalence of learning materials, etc.). This would provide a basis for

greater confidence in the meta-analysis conclusions and the resulting practical decisions that are made.

APPENDIX A – Coding of Study Characteristics

Table A1

Coded Study Characteristics Aligned with Conceptual Framework

Conceptual Framework Components (TPACK)	**Instructional Elements (IE)**	**Operationalization**
Context	**Analyze** Assess learners and identify what is to be learned.	
	Academic Level	
		Undergraduates
		Graduate
	Learners' Disciplinary Background	
		Interdisciplinary
		Same disciplines
	Student Gender Composition	
		Approximately Equal Number of Males & Females
		Majority Female
		Majority Male
	Course Disciplinary Area	
		Business
		Education or Social Sciences
		Physical, Natural, or Health Sciences
		Psychology
		Other
	Location	
		East
		North
		South
		West
		International

Table A1. (continued).

	Course Name	Business Statistics Criminal Justice Research Methods Introductory Social-Science/Social Statistics Introductory/Elementary Statistics Medical/Health Science Statistics Psychology Statistics Research Methods/Research Methods and Statistics
	Learning Goal	Develop statistical literacy, thinking or reasoning Learn statistical skills/concepts Understand the purpose (logic) or process of statistical investigations
	Learning Goal of Technology Use	Automation of calculations Collaboration and student involvement Investigation of real-life problems Simulation used as a teaching tool Visualization of concepts Multiple
	Cognitive Outcome Function of Technology	Information presentation Information seeking Knowledge integration knowledge organization
Content	Content	Descriptive Statistics, Hypothesis Testing Distributions, probability, centrality, randomness Data analysis/Inferential Statistics/Statistical Tests Multiple basic concepts (descriptive statistics, probability, sampling)

Table A1. (continued).

	Design How is it to be learned?	
Instructional Delivery Strategies	Treatment Duration	A semester or longer Less than a semester
	Mode of Instructional Delivery	Flipped/Hybrid/Blended/Distance Education FTF/Lab only FTF/Lecture only FTF/Lecture/Lab Online (All instruction online)
	Develop Produce or acquire instructional material.	
Technology	Media/Technology Type	Commercial statistical package Digital game Drill & practice or Web-based tutorial/Computer assisted learning LMS/CMS/Web-based course Multimedia/Presentation software Screencast tutorial/Vodcast Stand-alone or Web-based simulation/applet visualization tool Web information resource
	Technology Design	Institution hosted Propriety (commercial) Teacher/Researcher designed

Table A1. (continued).

	Cognitive Outcome Function of Technology	Information presentation Information seeking Knowledge integration knowledge organization
	Implementation Use the material and strategies to deliver instruction.	
Content and Pedagogy (C-P)	Learning Task	Assignments/Problem solving Lab exercises Multiple
Technology and Pedagogy (T-P)	Learner Engagement	Cooperative/collaborative/collective Individual Mixed (students work alone & in groups)
	Learner Control	Learner w/ materials only Learner with others (Teachers or peers)
	Scaffolding	Scaffolding Present None
	Feedback Type	Immediate Not immediate Both (immediate and not immediate) None
	Specificity of Feedback	Non-specific (provide correct or incorrect feedback only) Specific (provides feedback w/ detailed & specific response to behavior)

Table A1. (continued).

Technology and Content (T-C)	Technology function with concept	Computing (data analysis/diagnostics/ bootstrap)/graphing(distribution/outliers/models/centrality/ spread) Course management (collaboration) Data exploration Simulation(probability/variability) Multiple
	Evaluation Monitor and assessing the effectiveness of instruction.	
Pedagogy	Formative Assessment Measure	Homework assignment/practice questions/activities Quizzes/tests Multiple
	Summative Evaluation Type	Authentic assessment (e.g., assignment/project grade/presentation/demonstration/etc.) Non-authentic assessment (e.g., course grade/final/mid-term test/grade/exam/achievement test) Both
	Summative Assessment Measure	Another achievement test (e.g. teacher made exam/test/quiz) Standardized achievement/cognitive test Both (combined types of measures)

Note: The identification of variables and their operationalizations were defined based on the instructional design framework, literature review, and recommendations from Bates (2015), GAISE (2014), Garfield and Ben-zvi (2009), Garfield and Ben-zvi (2007), Harris, Mishra, and Koehler (2009), Means et al. (2009), Moore (1997), Sosa, Berger, Saw, & Mary (2011), Tishkovskaya and Lancaster (2012). Operationalizations with (e.g.) provide examples for the characteristics that will be identified, categorized, and assigned as levels of the variable based on findings in each primary study.

APPENDIX B – Threat to Validity Statements

Table B1.

Threat to Validity Statements

Validity Attribute	Validity Statement
Internal Validity	Pre-assessment of participants' technology acceptance/skills/competence?
	If pre-test was used, were pre-test and post-test versions different (Testing effect)?
	Did participants sampled represent either all low achieving or high achieving groups (Regression to Mean)?
	Is there any indication that participants who dropped out affected observed outcomes (Attrition)?
	Was the control group made aware of the treatment condition (Design contamination)?
	Data collection for experimental and control groups conducted at the same time or institution (History Effect)?
	Group Assignment (Selection Bias)
	Equivalence of groups established?
External Validity	Random Sampling of participants (Sampling Bias)?
	Participants relatively similar in age/gender/race?
Construct Validity	Score reliability or validity of outcome measurement instrument for current sample reported?
Implementation Fidelity	Equivalence of Curriculum Material
	Instructor Bias
	Implementation of Treatment and Control Conditions
Statistical Conclusion Validity	Confounds accounted for in analysis?
	Any indication of violations to any of assumptions (e.g. independence, normality, equal variance)?
	Any indication of a hierarchical/multilevel data structure (e.g. participants nested in different classrooms, teachers, or schools)?

APPENDIX C – Database and Keyword Search

Table C1

Database and Keyword Searches

Database Source	Keywords
Academic Search Premier; PsycINFO; Computers & Applied Sciences Complete; ER IC; Information Science & Technology Abstracts (ISTA);Newspaper Source; OpenDissertations	*statistics AND technology (OR all tech types) AND achievement AND introductory (OR higher education levels)*
Academic Search Premier ;ERIC; Information Science & Technology Abstracts (ISTA);Newspaper Source; OpenDissertations; PsycINFO	*statistics AND technology (OR all tech types) AND learning AND introductory (OR higher education levels)*
Academic Search Premier; PsycINFO; Computers & Applied Sciences Complete; ERIC; Information Science & Technology Abstracts (ISTA);Newspaper Source; OpenDissertations	*statistics AND technology (OR all tech types) AND cognitive AND introductory (OR higher education levels)*
Academic Search Premier; Computer Source; Computers & Applied Sciences Complete; ERIC; Information Science & Technology Abstracts (ISTA);Newspaper Source; OpenDissertations; PsycINFO	*statistics AND technology (OR all tech types) AND statistical thinking AND introductory (OR higher education levels)*
Academic Search Premier, Computers & Applied Sciences Complete, ERIC, Information Science & Technology Abstracts (ISTA), Newspaper Source, OpenDissertations, PsycINFO	*statistics AND technology (OR all tech types) AND statistical reasoning AND introductory (OR higher education levels)*

Table C1 (continued).

Database Source	Keywords
Academic Search Premier, Computers & Applied Sciences Complete, ERIC, Information Science & Technology Abstracts (ISTA), Newspaper Source, OpenDissertations, PsycINFO	*statistics AND technology (OR all tech types) AND statistical literacy AND introductory (OR higher education levels)*
Academic Search Premier, Computers & Applied Sciences Complete, ERIC, Information Science & Technology Abstracts (ISTA), Newspaper Source, OpenDissertations, PsycINFO	*statistics AND technology (OR all tech types) AND assessment AND introductory (OR higher education levels)*
Academic Search Premier, Computers & Applied Sciences Complete, ERIC, Information Science & Technology Abstracts (ISTA), Newspaper Source, OpenDissertations, PsycINFO	*statistics AND technology (OR all tech types) AND effectiveness AND introductory (OR higher education levels)*

APPENDIX D – Studies Excluded

Table D1

Explanations of Excluded Primary Studies

Did not meet criteria	**Reason**
(Christmann & Badgett, 1997)	No control/comparison of types of technologies
(Bell & Glen, 2008)	Not a quasi-experimental design
(Dempsey & Eck, 2003)	No control/comparison of types of technologies
(Devaney, 2010)	Non-cognitive outcomes measured
(Cherney, 2008)	No technology used as treatment condition
(Chow, Woodford, & Maes, 2011)	Insufficient statistical results provided to calculate effect size
(Cybinski & Selvanathan, 2005)	Insufficient statistical results provided to calculate effect size
(Debord, Aruguete, & Muhlig, 2004)	No comparison of technology vs no technology control on achievement alone
(Delcham & Sezer, 2010)	No comparison of technology vs no technology control on achievement alone
(delMas & Garfield, 1999)	No control/comparison of types of technologies
(Doğan, 2009)	No technology used
(Dunn, McDonald, & Loch, 2015)	Insufficient statistical data provided to calculate effect size
(Ferreira, Kataoka, & Karrer, 2014)	No comparison of technology vs no technology control

Table D1 (continued).

Did not meet criteria	Reason
(Frederickson & Reed, 1999)	Advanced psychology graduate student sample with prior undergraduate statistics experience
(Gopal, Salim, & Mohd Ayub, 2018)	High School students
(Green, 2007)	Anecdotal/Description of pedagogy using technology
(Grandzol, 2004)	Anecdotal/Description of pedagogy using technology
(Hagtvedt, Jones, & Jones, 2008)	Non-cognitive outcome measured
(Hammerman & Rubin, 2004)	Middle-school students and high school teachers, qualitative
Hodgson, Pang (2012)	No comparison control group
Hurlburt (2001)	No comparison control group
(Lajoie, 1997)	Anecdotal, qualitative
(Lane & Tang, 2000)	Qualitative study
(Mcgowan & Gunderson, 2010)	Single group design
(Mclaughlin & Kang, 2017)	Single group pre-post
(Messecar, Van Son, & O'Meara, 2003)	No control/comparison of types of technologies
(Jamie D Mills, 2002)	Review of type of technology
(Mills & Johnson, 2004)	Anecdotal, qualitative
(Novak, 2012)	Technology used in both treatment and control
(Palocsay & Stevens, 2008)	Insufficient statistical data provided to calculate effect size
(Petty, 2010)	Anecdotal/qualitative
(Phillips & Phillips, 2016)	No treatment vs. control
(Makar & Sousa, 2014)	Anecdotal, qualitative
(Porter, Griffiths, & Hedberg, 2003)	Anecdotal, qualitative
(Prodromou, 2014)	Anecdotal, qualitative
(Prodromou, 2015)	Anecdotal, qualitative

Table D1 (continued).

Did not meet criteria	Reason
(Quilter, 2001)	Single-group pretest-posttest design
(Raffle & Brooks, 2005)	No control group
(Ramesh, 2011)	Anecdotal, qualitative
(Ray, Leeper, & Amini, 2014)	No technology, cooperative learning only
(Reaburn, 2014)	No tech treatment vs control conditions
(Roberts, 2007)	No control group
(Sabbag & Zieffler, 2015)	Focus on psychometric analysis of Goals-2 instrument
(Schuyten & Thas, 2007)	Anecdotal, qualitative
(Stephenson, 2001)	Insufficient statistical data provided to calculate effect size
(Suanpang, Petocz, & Kalceff, 2004)	Non-cognitive outcome measured
(Summers, Waigandt, & Whittaker, 2005)	Insufficient statistical data provided to calculate effect size
(Symanzik & Vukasinovic, 2003)	Anecdotal (description of implementation of technology-enhanced course)
(Timmerman & Kruepke, 2006)	Meta-analysis study on CAI in general education
(Tu & Snyder, 2017)	Single group design, non-experimental
(Velleman & Moore, 1996)	Anecdotal
(Wessa, Rycker, & Holliday, 2011)	Comparison of two types of VLE technologies
(West & Ogden, 1998)	Anecdotal (description/example of implementation technology)
(Wit, 2003)	Qualitative
(Xu, Zhang, Su, Cui, & Qi, 2014)	Single group design, non-experimental
Wender, K. F., Muehlboeck, J. (2003)	Insufficient statistical data provided to calculate effect size

APPENDIX E – Cohen's Kappa Computation

Table E1

Results of Inter-Rater Reliability Computation

Characteristics	**Cohen's Kappa**
Study	
Academic Level	0.68
Learners' Academic Backgrounds	0.88
Subject/Disciplinary Area	0.90
Treatment Duration	0.84
Learning Goal	0.88
Learning Objective(s)	0.93
Mode of instruction/Delivery Format	0.67
Media/Technology Type	0.79
Technology Design	0.67
Learning Goal Function of Technology	0.79
Cognitive Outcome Function of Technology	0.83
Learning Task	0.78
Learner Engagement	0.76
Learner Control	0.72
Scaffolding	0.80
Feedback Type	0.77
Specificity of Feedback	0.68
Technology Function with Concept	0.81
Formative Assessment Measure	0.69
Summative Assessment Measure	0.80
Summative Evaluation Type	0.75
Average	**0.78**
Methodological Characteristics	
Research Design	1.00
Instructor Bias	0.80
Material Equivalence	0.88
Implementation of Treatment & Control	0.78

Table E1 (continued).

Characteristics	Cohen's Kapa
Methodological	
Description of ID process	.78
Reported psychometric properties of outcome measurement instrument	.00
Type of outcome measure used	.75
Funded research	.00
Mentioned potential confounds (Y/N)	.75
Average	**.86**
Study Quality	
Pre-assessment of participants' technology acceptance/skills/competence?	.83
If pre-test was used, were pre-test and post-test versions different or different forms used (i.e. parallel forms) (Testing effect)?	.80
Did participants sampled represent either all low achieving or high achieving groups (Regression to Mean)?	.73
Was attrition present or any indication that participants who dropped out could have affected observed outcomes (Attrition)?	.75
Was the control group made aware of the treatment condition (Design contamination)?	.87
Confounds Addressed?	.82
Data collection for experimental and control groups conducted at the same time or institution (History Effect)?	.81
Group Assignment (Selection Bias)	.81
Equivalence of groups established?	.64
Random Sampling of participants (Sampling Bias)?	.89
Demographic characteristics (e.g. age, gender, race):	.71
Score reliability or validity of outcome measurement instrument for current sample reported?	.80
Equivalence of Curriculum Materia	.80
Instructor Bias	.00

Characteristics	Cohen's Kapa
Any indication of violations to any of assumptions (e.g. independence, normality, equal variance)?	.73
Any indication of a hierarchical/multilevel data structure (e.g. participants nested in different classrooms, teachers, or schools)?	.90
Average	**.81**

APPENDIX F – Effect Size Computation

Table F1

Methods of Effects Size Computation

Author	Test Used	Statistic Used to calculate the ES	Calculation
Aberson, et al.	ANCOVA(One-Factor)	p-value	Estimate from partial inferential statistics (e.g. p-value)
Aberson, et al.	ANCOVA(One-Factor)	p-value	Estimate from partial inferential statistics (e.g. p-value)
Arena, D., & Schwartz, D.	T-test	t-test	ES reported by authors (only used when no other information is available)
Arena, D., & Schwartz, D.	T-test	t-test	ES reported by authors (only used when no other information is available)
Basturk, R.	T-test	t-test	Calculated with inferential statistics
Basturk, R.	T-test	t-test	Calculated with inferential statistics
Benedict, J. O., & Anderton, J. B.	T-test	Means, SD	Calculated with descriptive statistics

Table F1 (continued).

Author	Test Used	Statistic Used to calculate the ES	Calculation
Burruss, G. W., & Furlow, M. H.	Fischer's Exact	p-value	Estimate from partial inferential statistics (e.g. p-value)
Burruss, G. W., & Furlow, M. H.	Mann-Whitney U	Means, SD	Calculated with descriptive statistics
Burruss, G. W., & Furlow, M. H.	Mann-Whitney U	Means, SD	Calculated with descriptive statistics
Burruss, G. W., & Furlow, M. H.	Mann-Whitney U	Means, SD	Calculated with descriptive statistics
Ciftci, S. K., Karadag, E., & Akdal, P.	ANCOVA(One-Factor)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Dinov, I. D., Sanchez, J., & Christou, N.	T-test	t-test	Calculated with inferential statistics
Frederickson, N., Reed, P., & Clifford, V.	F-Test(Multi-factor ANOVA)	Means, SD	Calculated with descriptive statistics
Gonzalez, G. M., & Birch, M. A.	F-Test(Multi-factor ANOVA)	Means, SD	Calculated with descriptive statistics

Table F1 (continued).

Author	Test Used	Statistic Used to Calculate the ES	Calculation
Gonzalez, G. M., & Birch, M. A.	F-Test (Multi-factor ANOVA)	Means, SD	Calculated with descriptive statistics
High, R. V.	T-test	Means, SD	Calculated with descriptive statistics
Hilton, S. C., & Christensen, H. B.	Linear Mixed Model/ F-test (Fixed Factor – other factors controlled for)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Hilton, S. C., & Christensen, H. B.	Linear Mixed Model/ F-test (Fixed Factor – other factors controlled for)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Hilton, S. C., & Christensen, H. B.	Linear Mixed Model/ F-test (Fixed Factor – other factors controlled for)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Hilton, S. C., & Christensen, H. B.	Linear Mixed Model/ F-test (Fixed Factor – other factors controlled for)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Jones, E. R.	Repeated Measures ANOVA	p-value	Estimate from partial inferential statistics (e.g. P-value)

Table F1 (continued).

Author	Test Used	Statistic Used to calculate the ES	Calculation
Lane, J. L., & Aleksic, M.	F-Test (One Factor ANOVA)	F-test	Calculated with inferential statistics
Lane, J. L., & Aleksic, M.	F-Test (One Factor ANOVA)	F-test	Calculated with inferential statistics
Lane, J. L., & Aleksic, M.	F-Test (One Factor ANOVA)	F-test	Calculated with inferential statistics
Larwin, K. H., & Larwin, D. A.	F-Test (One Factor ANOVA)	Means, SD	Calculated with descriptive statistics
Larwin, K. H., & Larwin, D. A.	T-test	Means, SD	Calculated with descriptive statistics
Lloyd, S. A., & Robertson, C. L.	T-test	Means, SD	Calculated with descriptive statistics
Lloyd, S. A., & Robertson, C. L.	T-test	Means, SD	Calculated with descriptive statistics
Lu, F., & Lemonde, M.	T-test	Means, SD	Calculated with descriptive statistics
Maurer, K., & Lock, D.	ANCOVA(Multiple Factors)	F-test	Calculated with inferential statistics

Table F1 (continued).

Author	Test Used	Statistic Used to Calculate the ES	Calculation
McLaren, C. H.	Chi Square	p-value	Estimate from partial inferential statistics (e.g. P-value)
Milic, et al.	F-Test (One Factor ANOVA)	Means, SD	Calculated with inferential statistics
Mills, J. D.	T-test	p-value	Estimate from partial inferential statistics (e.g. P-value)
Mills, J. D.	F-Test (One Factor ANOVA)	p-value	Estimate from partial inferential statistics (e.g. P-value)
Morris, E.	T-test	Means, SD	Calculated with descriptive statistics
Peterson, D. J.	T-test	Means, SD	Calculated with descriptive statistics
Petta, N. A.	T-test	Means, SD	Calculated with descriptive statistics

Table F1 (continued).

Author	Test Used	Statistic Used to Calculate the ES	Calculation
Ragasa, C. Y.	ANCOVA(One-Factor)	Means, SD	ES reported by authors (only used when no other information is available)
Smith, T.	T-test	t-test	Calculated with inferential statistics
Spinelli, M. A.	T-test	Means, SD	Calculated with descriptive statistics
Tintle et al.	Paired T-test	Means, SD	Calculated with descriptive statistics
Utts et al.	ANCOVA(One-Factor)	Means, SD	Calculated with descriptive statistics
Wang, A. Y., & Newlin, M. H.	T-test	Means, SD	Calculated with descriptive statistics
Wang, X.	T-test	Means, SD	Calculated with descriptive statistics
Wilmoth, J., & Wybraniec, J.	F-Test	Means, SD	Calculated with descriptive statistics

APPENDIX G – Leave-One-Out Output

Table G1

Leave-One-Out Analysis Output for K = 32 Studies

	estimate	zval	pval	ci.lb	ci.ub	Q	Qp	tau2	I2	H2
1	1.227	2.257	0.024	1.027	1.466	307.132	0.000	0.201	93.932	16.479
2	1.279	2.805	0.005	1.077	1.519	297.715	0.000	0.182	93.246	14.807
3	1.241	2.340	0.019	1.036	1.487	309.670	0.000	0.208	94.098	16.943
4	1.206	2.126	0.034	1.015	1.433	257.789	0.000	0.183	93.110	14.513
5	1.236	2.286	0.022	1.031	1.482	306.850	0.000	0.208	94.033	16.759
6	1.260	2.515	0.012	1.052	1.510	307.770	0.000	0.204	93.761	16.027
7	1.230	2.248	0.025	1.027	1.473	304.764	0.000	0.205	93.979	16.608
8	1.263	2.551	0.011	1.056	1.511	307.312	0.000	0.202	93.809	16.152
9	1.244	2.393	0.017	1.040	1.488	310.289	0.000	0.205	94.065	16.849
10	1.209	2.156	0.031	1.017	1.437	295.827	0.000	0.186	93.431	15.224
11	1.263	2.557	0.011	1.056	1.510	308.279	0.000	0.201	93.878	16.334
12	1.254	2.438	0.015	1.045	1.505	244.940	0.000	0.209	88.306	8.551
13	1.235	2.291	0.022	1.031	1.480	308.142	0.000	0.207	94.061	16.838
14	1.229	2.228	0.026	1.025	1.473	207.739	0.000	0.206	91.368	11.585
15	1.211	2.154	0.031	1.017	1.442	291.125	0.000	0.189	93.496	15.375
16	1.210	2.159	0.031	1.018	1.439	296.523	0.000	0.186	93.458	15.285
17	1.271	2.679	0.007	1.066	1.515	305.322	0.000	0.192	93.640	15.722
18	1.240	2.317	0.020	1.034	1.486	308.677	0.000	0.209	94.072	16.870
19	1.243	2.338	0.019	1.036	1.491	308.740	0.000	0.210	94.003	16.675
20	1.246	2.397	0.017	1.041	1.491	310.310	0.000	0.206	94.079	16.889
21	1.219	2.194	0.028	1.021	1.455	301.383	0.000	0.196	93.761	16.029
22	1.253	2.457	0.014	1.047	1.499	310.207	0.000	0.205	94.049	16.805
23	1.225	2.233	0.026	1.025	1.464	305.524	0.000	0.201	93.910	16.420
24	1.247	2.402	0.016	1.041	1.493	310.325	0.000	0.207	94.086	16.908
25	1.225	2.234	0.026	1.025	1.464	305.530	0.000	0.201	93.913	16.428
26	1.241	2.341	0.019	1.036	1.488	309.666	0.000	0.208	94.098	16.944
27	1.275	2.735	0.006	1.071	1.518	300.908	0.000	0.188	93.444	15.253
28	1.263	2.558	0.011	1.056	1.511	308.022	0.000	0.201	93.862	16.291
29	1.269	2.630	0.009	1.063	1.516	288.920	0.000	0.197	93.083	14.457
30	1.270	2.673	0.008	1.066	1.514	305.801	0.000	0.193	93.660	15.774
31	1.242	2.349	0.019	1.036	1.487	309.858	0.000	0.208	94.097	16.941
32	1.242	2.327	0.020	1.035	1.490	307.888	0.000	0.210	93.965	16.570

APPENDIX H –Influence Diagnostic Measures

Table H1

Output for Influence Diagnostics

	Std. Residual	Cook's Distance
Aberson, C. L., et al.1	0.91	0.12
Aberson, C. L., et al.2	-1.41	0.60
Arena, D., & Schwartz, D.	0.83	0.17
Basturk, R.	0.16	0.41
Benedict, J. O., & Anderton, J. B.	0.17	8.29e -3
Burruss, G. W., & Furlow, M. H.	0.12	7.44e -3
Ciftci, S. K., Karadag, E…	0.25	0.06
Dinov, I. D., Sanchez, J…	-0.61	0.06
Frederickson, N., Reed, P…	-0.15	4.35e -3
Gonzalez, G. M., & Birch, M. A.	0.42	0.65
High, R. V.	-0.05	1.82e -3
Hilton, S. C., & Christensen, H. B.	-0.63	0.18
Jones, E. R.	0.40	0.05
Lane, J. L., & Aleksic, M.	0.13	6.59e -3
Larwin, K. H., & Larwin, D. A.	0.97	0.15
Lloyd, S. A., & Robertson, C. L.	1.59	1.68
Lu, F., & Lemonde, M.	-0.78	0.20
Maurer, K., & Lock, D.	-0.51	0.04
McLaren, C. H.	-0.73	0.10
Milic, N. M., et al.	-0.18	0.01
Mills, J. D.	2.05	1.11
Morris, E.	-0.81	0.40
Peterson, D. J.	0.42	0.14
Petta, N. A.	-0.70	0.08
Ragasa, C. Y.	0.41	0.02
Smith, T.	-0.64	0.26
Spinelli, M. A.	-2.23	1.06
Tintle, N., et al.	-0.36	0.07
Utts, J., et al.	-0.62	0.17
Wang, A. Y., & Newlin, M. H.	-1.87	0.87
Wang, X.	1.16	0.49
Wilmoth, J., & Wybraniec, J.	-0.18	0.02

REFERENCES

Articles with an asterisk represent those used in the meta-analysis.

*Aberson, C. L., Berger, D. E., Healy, M. R., Kyle, D. J., & Romero, V. L. (2000). Evaluation of an interactive tutorial for teaching the central limit theorem. *Teaching in Psychology*, *27*(4), 289–291.

*Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2003). Evaluation of an interactive tutorial for teaching hypothesis testing concepts. *Teaching of Psychology*, *30*(1), 75–78. https://doi.org/10.1207/S15328023TOP3001_12

Archer, K., Savage, R., Sanghera-Sidhu, S., Wood, E., Gottardo, A., & Chen, V. (2014). Examining the effectiveness of technology use in classrooms: A tertiary meta-analysis. *Computers and Education*, *78*, 140–149. https://doi.org/10.1016/j.compedu.2014.06.001

*Arena, D., & Schwartz, D. (2014). Experience and explanation: Using video games to prepare students for formal instruction in statistics. *Journal of Science Education & Technology*, *23*(4), 538–548. Retrieved from http://10.0.3.239/s10956-013-9483-3

*Basturk, R. (2005). The effectiveness of computer-assisted instruction in teaching introductory statistics. *Educational Technology & Society*, *8*(2), 170–178.

*Benedict, J. O., & Anderton, J. B. (2004). Applying the just-in-time teaching approach to teaching statistics. *Teaching of Psychology*, *31*(3), 197–199.

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., & Surkes, M. A. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research*, *79*, 1243–1289. https://doi.org/10.3102/0034654309333844

Bernard, R. M., Borokhovski, E., Schmid, R. F., & Tamim, R. M. (2014). An exploration of bias in meta-analysis: The case of technology integration research in higher education. *Journal of Computing in Higher Education*, *26*(3), 183–209. https://doi.org/10.1007/s12528-014-9084-z

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education*, *26*, 87–122. https://doi.org/10.1007/s12528-013-9077-3

*Burruss, G. W., & Furlow, M. H. (2007). Teaching statistics visually: A quasi-experimental evaluation of teaching chi-square through computer tutorials. *Journal of Criminal Justice Education*, *18*(2), 209–230. https://doi.org/10.1080/10511250701383343

Borenstein, M., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2017). *Introduction to meta-analysis*. West Sussex: John Wiley & Sons, Ltd.

Borokhovski, E., Bernard, R. M., Tamim, R. M., Schmid, R. F., & Sokolovskaya, A. (2016). Technology-supported student interaction in post-secondary education: A meta-analysis of designed versus contextual treatments. *Computers & Education*, *96*, 15–28. https://doi.org/10.1016/J.COMPEDU.2015.11.004

Boyer, N., Beard, J., Holt, L., Larsen, J., Piskurich, J., & Piskurich, G. (2013). Fostering technology and self-direction: The impact on adults in education, business, and everyday life. *International Journal of Self-Directed Learning*, *10*(2), 1–20. Retrieved from www.sdlglobal.com

Card, N. A. (2012). *Applied meta-Analysis for social science research*. New York: The Guilford Press.

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), 1–26. Retrieved from file:///C:/Users/jacquel7/Downloads/eScholarship UC item 8sd2t4rr.pdf

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, 29(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

*Ciftci, S. K., Karadag, E., & Akdal, P. (2014). Instruction of statistics via computer-based tools: effects on statistics' anxiety, attitude, and achievement. *Journal of Educational Computing Research*, *50*(1), 119–133. Retrieved from http://10.0.8.142/EC.50.1.f

Clark, R. E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing questioning the meta-analyses of computer-based instruction research . *Educational Communications and Technology Journal*, *33*(4), 249–262. https://doi.org/10.1007/BF02769362

Clark, R. E. (2001). *Learning from instructional media: Arguments, analysis, and evidence.* Greenwich: Information Age Publishing.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development*, *42*(2), 21–29. https://doi.org/http://dx.doi.org/10.1007/BF02299088

Cobb, G. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change suggestions for curricular action* (pp. 3–43). Washington DC: The Mathematical Association of America.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375–395). https://doi.org/10.1007/1-4020-2278-6_4

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). New York: Academic Press.

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

Cooper, H. (2017). *Research synthesis and meta-analysis* (5th ed.). Thousand Oaks, CA: Sage Publications Inc.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–16). New York: Russel Sage Foundation.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russel Sage Foundation.

Denson, N., & Seltzer, M. H. (2010). Meta-analysis in higher education: An illustrative example using hierarchical linear modeling. *Research in Higher Education*, *52*(3), 215–244. https://doi.org/10.1007/s11162-010-9196-x

*Dinov, I. D., Sanchez, J., & Christou, N. (2008). Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses. *Computers and Education*, *50*(1), 284–300. https://doi.org/10.1016/j.compedu.2006.06.003

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical & Statistical Psychology*, *63*(3), 665–694. https://doi.org/10.1348/000711010×502733

Francis, D. C., Hudson, R., & Vesperman, C. (2014). Comparing technology-supported teacher education curricular models for enhancing statistical content knowledge. *Interdisciplinary Journal of Problem-Based Learning Volume*, *8*(1).

Franklin, C., & Garfield, J. (2006). Guidelines for statistics education endorsed by ASA Board of Directors. *Amstat News (Education)*, (348).

*Frederickson, N., Reed, P., & Clifford, V. (2005). Evaluating web-supported learning versus lecture-based teaching: Quantitative and qualitative perspectives. *Higher Education*, *50*(4), 645–664. https://doi.org/10.1007/s10734-004-6370-0

Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., & Russell, J. D. (2005). *Principles of instructional design* (5th ed.). Belmont: Thomson/Wadsworth.

Garfield, J. B., Ben-zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* https://doi.org/https://doi.org/10.1007/978-1-4020-8383-9

Garfield, J., & Ben-zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Revie*, *75*(3), 372–396. https://doi.org/10.1111/j.l751-5823.2007.00029.x

Garfield, J., Chance, B., Poly-San, C., & Obispo, L. (1999). Assessment in statistics education: Issues and challenges. *International Statistical Review, 67*(1), 1–12. https://doi.org/10.2307/1403562

Garfield, J., DelMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*, 883–898. https://doi.org/10.1007/s11858-012-0447-5

Gebre, E., Saroyan, A., & Bracewell, R. (2014). Students' engagement in technology rich classrooms and its relationship to professors' conceptions of effective teaching. *British Journal of Educational Technology*, *45*(1), 83–96. https://doi.org/10.1111/bjet.12001

*Gonzalez, G. M., & Birch, M. A. (2000). Evaluating the instructional efficacy of computer-mediated interactive multimedia: Comparing three elementary statistics tutorial modules. *Journal of Educational Computing Research, 22*(4), 411–436. https://doi.org/10.2190/X8PQ-K0GQ-T2DR-XY1A

González, J. A., Jover, L., Cobo, E., & Muñoz, P. (2010). A web-based learning tool improves student performance in statistics : A randomized masked trial. *Computers & Education Journal*, *55*, 704–713. https://doi.org/10.1016/j.compedu.2010.03.003

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). Doing meta-Analysis in R: A hands-on Guide. DOI: 10.5281/zenodo.2551803. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Harris, J., Mishra, P., & Koehler, M. (2009). Teachers' technological pedagogical content knowledge and learning activity types: Curriculum-based technology integration

reframed. *Journal of Research on Technology in Education*, *41*(4), 393–416. https://doi.org/10.1207/s15326985ep2803_7

Hassad, R. A. (2009). Reform-oriented teaching of introductory statistics in the health, social and behavioral sciences - historical context and rationale. *World Academy of Science, Engineering and Technology*, *40*, 398–403. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-78651555277&partnerID=40&md5=9ac16f172e72dd0ec05947ac9a15e4a5

Hassad, R. A. (2009). Development and validation of a teaching practice scale (TISS) for instructors of introductory statistics at the college level. *International Association for Statistical Education - International Statistical Institute, 2009*, 1–8. https://doi.org/10.5038/1936-4660.4.2.7

Hassad, R. A. (2011). Constructivist and behaviorist approaches: Development and initial evaluation of a teaching practice scale for introductory statistics at the college level. *Numeracy*, *4*(2). https://doi.org/http://dx.doi.org/10.5038/1936-4660.4.2.7

Hassad, R. A. (2014). The status of reform in statistics education: A focus on the introductory course. *ICOTS9 Contributed Paper Proceedings*, *9*, 1–4.

Hedges, L. V, & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Hedges L.V., & Vevea, J.L. (1998). "Fixed-and random-effects models in meta-analysis." *Psychological Methods*, 3(4), 486–504.

Higgins, J. P. T., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis." *Statistics in Medicine, 21* (11). Wiley Online Library: 1539–58.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557-560.

*High, R. V. (1998). *Some variables in relation to students' choice of statistics classes: Traditional versus computer-supported instruction*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED427762&site=ehost-live

*Hilton, S. C., & Christensen, H. B. (2002). Evaluating the impact of multimedia lectures on student learning and attitudes. *Developing a Statistically Literate Society. Proceedings of the Sixth International Conference on Teaching of Statistics*, 1–6.

Hsu, Y. (2003). *The effectiveness of computer-assisted instruction in statistics education: A meta- analysis.* The University of Arizona.

Hunter, J., & Schmidt, F. L. (2004). *Methods of meta-Analysis: Correcting error and bias in research findings.* (2nd ed.; T. Hunter, Ed.). Thousand Oaks, CA: Sage:

Iiyoshi, T., Hannafin, M. J., & Wang, F. (2005). Cognitive tools and student-centered learning: Rethinking tools, functions and applications. *Educational Media International*, *42*(4), 281–296.

*Jones, E. R. (1999). A comparison of an all web-based class to a traditional class. *Paper Presented at the Annual Conference of the Society for Information Technology & Teacher Education*. San Antonio, TX.

Kalaian, S. A., & Kasim, R. M. (2014). A meta-analytic review of studies of the effectiveness of small-group learning methods on statistics achievement. *Journal of Statistics Education*, *22*(1), 1–20.

Kennewell, S. (2001). Using affordances and constraints to evaluate the use of information and communications technology in teaching and learning. *Journal of Information Technology for Teacher Education*, *10*(1–2), 101–116. https://doi.org/10.1080/14759390100200105

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. *Educational Technology Research and Development*, *42*(2), 7–19. https://doi.org/10.1007/BF02299087

Kulik, C.-L. C., Kulik, J. A., & Shwalb, B. J. (1986). The effectiveness of computer-based adult education: A meta-analysis. *Journal of Educational Computing Research*, *2*(2), 235–252. https://doi.org/10.2190/02HM-XCWG-Q1VY-5EMQ

Kulik, C. L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, *7*(1–2), 75–94. https://doi.org/10.1016/0747-5632(91)90030-5

Kulik, J. A., Kulik, C. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Educational Research*, *50*(4), 525–544.

Lajoie, S. P. (1997). Technologies for assessing and extending statistical learning. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (Vol. 90, pp. 179–190). Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/assessbkref.

*Lane, J. L., & Aleksic, M. (1998). Transforming elementary statistics to enhance student learning. *Paper Presented at the Annual Meeting of the American Educational Research Association*, 19. https://doi.org/10.1158/1535-7163.mct-16-0142

Larwin, K., & Larwin, D. (2011). A meta-analysis examining the impact of computer-assisted instruction on postsecondary statistics education: 40 years of research. *Journal of Research on Technology in Education*, *43*(3), 253–278. https://doi.org/10.1080/15391523.2011.10782572

*Larwin, K. H., & Larwin, D. A. (2011). Evaluating the use of random distribution theory to introduce statistical inference concepts to business students. *Journal of Education for Business*, *86*(1), 1–9. https://doi.org/10.1080/08832321003604920

Lin, L., Chu, H., & Hodges, J. S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics*, 73(1), 156–166. doi:10.1111/biom.12543

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. SAGE publications, Inc.

*Lloyd, S. A., & Robertson, C. L. (2012). Screencast tutorials enhance student learning of statistics. *Teaching of Psychology*, *39*(1), 67–71. Retrieved from http://10.0.4.153/0098628311430640

Lou, Y., Abrami, P., & D'Apollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of Educational Research*, *71*(3), 449–521. Retrieved from http://rer.sagepub.com/content/71/3/449.short

Lowyck, J. (2014). Bridging learning theories and technology-enhanced environments: A critical appraisal of its history. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology: Fourth Edition* (4th ed., pp. 3–20). https://doi.org/10.1007/978-1-4614-3185-5_1

Lin, L., ,& Chu H. (2020). Altmeta: Alternative meta-analysis methods. R package version 2.2. Retrieved from https://CRAN.R-project.org/package=altmeta.

*Lu, F., & Lemonde, M. (2013). A comparison of online versus face-to-face teaching delivery in statistics instruction for undergraduate health science students. *Advances in Health Sciences Education*, *18*(5), 963–973. https://doi.org/10.1007/s10459-012-9435-3

Makar, I. K., & Sousa, B. De. (2014). How a curriculum may develop technological statistical knowledge: A case of teachers examining relationships among variables using Fathom. *ICOTS9 Contributed Paper Proceedings*, *9*, 1–6.

*Maurer, K., & Lock, D. (2016). Comparison of learning outcomes for simulation-based and traditional inference curricula in a designed educational experiment. *Technology Innovations in Statistics Education*, *9*(1), 1–20. https://doi.org/10.5811/westjem.2011.5.6700

Mcgowan, H. M., & Gunderson, B. K. (2010). A randomized experiment exploring how certain features of clicker use effect undergraduate students' engagement and learning in statistics. *Technology Innovations in Statistics Education*, *4*(1).

Mcgrath, A. L. (2014). Content, affective, and behavioral challenges to learning: Students' experiences learning statistics. *International Journal for the Scholarship of Teaching and Learning*, *8*(2). Retrieved from http://digitalcommons.georgiasouthern.edu/ij-sotl%5Cnhttp://digitalcommons.georgiasouthern.edu/ij-sotl/vol8/iss2/6

McGraw, J. B., & Chandler, J. L. (2015). Flipping the biostatistics classroom, with a twist. *Bulletin of the Ecological Society of America*, *96*(April), 375–384. Retrieved from https://doi.org/10.1890/0012-9623-96.2.37

McGuinness, L. A. (2019). robvis: An R package and web application for visualising risk-of-bias assessments. Retrieved from https://github.com/mcguinlu/robvis

*McLaren, C. H. (2004). A comparison of student persistence and performance in online and classroom business statistics experiences. *Decision Sciences Journal of Innovative Education*, Vol. 2, pp. 1–10. https://doi.org/10.1111/j.0011-7315.2004.00015.x

Mclaughlin, J. E., & Kang, I. (2017). A flipped classroom model for a biostatistics short course. *Statistics Education Research Journal*, *16*(2), 441–453.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. In *U.S. Department of Education*. https://doi.org/10.1016/j.chb.2005.10.002

*Milic, N. M., Trajkovic, G. Z., Bukumiric, Z. M., Cirkovic, A., Nikolic, I. M., Milin, J. S., … Stanisavljevic, D. M. (2016). Improving education in medical statistics: Implementing a blended learning model in the existing curriculum. *PLoS ONE*, *11*(2), 1–11. https://doi.org/10.1371/journal.pone.0148882

*Mills, J. D. (2004). Learning abstract statistics concepts using simulation. *Educational Research Qurterly*, (1996), 18–34.

Mishra, P., Koehler, M. J., & Bragg, W. H. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

*Morris, E. (2001). The design and evaluation of Link: A computer-based learning system for correlation. *British Journal of Educational Technology*, *32*(1), 39–52.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods, 7(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Morrison, G. R., & Ross, Steven, M. (2014). Research based instructional perspectives. In Michael J. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications technology* (4th ed., pp. 31–38). New York: Springer Science & Business Media.

Ooms, A., & Garfield, J. (2008). A model to evaluate online educational resources in statistics. *Technology Innovations in Statistics Education*, *2*(1).

Pearl, D. K., Garfield, J. B., DelMas, R. C., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting research to practice in a culture of assessment for introductory college-level statistics*. Retrieved from http:// www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf

*Peterson, D. J. (2016). The flipped classroom improves student achievement and course satisfaction in a statistics course. *Teaching of Psychology*, *43*(1), 10–15. Retrieved from http://10.0.4.153/0098628315620063

*Petta, N. A. (1999). *The impact of a web-based class management system on student performance and attitudes in a quantitative statistics class*. University of Nebraska.

Phillips, L., & Phillips, M. (2016). Improved student outcomes in a flipped statistics course. *Administrative Issues Journal: Connecting Education, Practice, and Research*, *6*(1), 88–98. https://doi.org/10.5929/2016.6.1.5

Price, S., & Oliver, M. (2007a). A framework for conceptualising the impact of technology on teaching and learning. *Educational Technology & Society*, *10*(1), 16–27. https://doi.org/10.1080/03054985.2015.1035703

Price, S., & Oliver, M. (2007b). Technology and change in educational practice (Guest Editorial). *Educational Technology & Society*, *10*, 14–15.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from https://www.R-project.org/

*Ragasa, C. Y. (2008). A comparison of computer-assisted instruction and the traditional method of teaching basic statistics. *Journal of Statistics Education*, *16*(1). https://doi.org/10.1080/10691898.2008.11889556

Ramesh, N. (2011). *Use of technology to enhance teaching and learning in mathematics and statistics*. *MSOR Connections*, *11*(1), 34–36.

Robinson, D., Lee, M., & Soutar, A. (2009). Using technology to direct learning. *Active Learning in Higher Education*, *10*(1), 71–83. https://doi.org/10.1177/1469787408100196

Roblyer, M. D. (2005). Educational technology research that makes a difference: Series introduction. *Educational Researcher*, *5*(2), 192–201. Retrieved from http://www.citejournal.org/vol5/iss2/seminal/article1.cfm

Ronau, R. N., Niess, M. N., Browning, C. A., Pugalee, D., Driskell, S. O., & Harrington, R. (2008). Framing research on technology and student learning in the content areas: Mathematics. In A. Thompson, L. Bell, & L. Schrum (Eds.), *Framing research on technology and student learning in the content areas: Implications for educators* (pp. 13–31). Charlotte: Information Age Publishing.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-Analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*(1), 59–82.

Roseth, C., Akcaoglu, M., & Zellner, A. (2013). Blending synchronous face-to-face and computer-supported cooperative learning in a hybrid doctoral seminar. *TechTrends*, *57*(3), 54–59. https://doi.org/10.1007/s11528-013-0663-z

Ross, S. M., & Morrison, J. R. (2014). Measuring meaningful outcomes in consequential contexts : Searching for a happy medium in educational technology research (Phase II). *Journal of Computing in Higher Education*, *26*, 4–21. https://doi.org/10.1007/s12528-013-9074-6

Schenker, J. D. (2007). *The effectiveness of technology use in statistics instruction in higher education: A meta-analysis using hierarchical linear modeling*. Kent State University. Retrieved from http://lynx.lib.usm.edu/login?url=https://search.proquest.com/docview/304835263?accountid=13946

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., … Woods, J. (2014). The effects of technology use in postsecondary

education: A meta-analysis of classroom applications. *Computers & Education*, *72*, 271–291. https://doi.org/10.1016/j.compedu.2013.11.002

Schrum, L., Thompson, A., Maddux, C., Sprague, D., Bull, G., & Bell, L. (2007). Editorial: Research on the effectiveness of technology in schools: The roles of pedagogy and content. *Contemporary Issues in Technology and Teacher Education*, *7*(1), 456–460. Retrieved from http://www.editlib.org/p/26278/

Schwier, R., & Seaton, J. (2013). Comparison of participation patterns in selected formal, non-formal, and informal online learning environments. *Canadian Journal of Learning & Technology*, *39*(1), 1–15. Retrieved from http://lynx.lib.usm.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=87378835&site=ehost-live

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, Massachusetts: Houghton Mifflin Company.

Sharma, S. (2017). Open review of educational research definitions and models of statistical literacy : A literature review. *Open Review of Educational Research*, *4*(1), 118–133. https://doi.org/10.1080/23265507.2017.1354313

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analytic and traditional reviews. *Journal of Clinical Epidemiology*, *48*(1), 9–18. https://doi.org/10.1016/0895-4356(94)00097-A

*Smith, T. (2017). Gamified modules for an introductory statistics course and their impact on attitudes and learning. *Simulation and Gaming*, *48*(6), 832–854. https://doi.org/10.1177/1046878117731888

Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research*, *81*(1), 97–128. https://doi.org/10.3102/0034654310378174

Spector, Michael J., Merrill, M. D., Elen, J., & Bishop, M. J. (2014). *Handbook of research on educational communications and technology* (4th ed.; M. J. Spencer, M. D. Merrill, J. Elen, & M. J. Bishop, Eds.). New York: Springer Science & Business Media B.V.

Spector, Michael J., Merrill, M. D., van Merrienboer, J. J. G., & Driscoll, M. P. (2008). *Handbook of research on educational communications technology* (M. J. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll, Eds.). New York: Routledge Taylor & Francis.

*Spinelli, M. A. (2001). The use of technology in teaching business statistics. *Journal of Education for Business*. https://doi.org/10.1080/08832320109599669

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 125–138). New York: Russel Sage Foundation.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research*, *81*(1), 4–28. https://doi.org/10.3102/0034654310393361

The Cochrane Collaboration (2011). Chapter 8: Assessing risk of bias in included studies. In: Higgins J. P. T., Altman, D. G., Sterne, J. A. C. (editors). *Cochrane Handbook*

for Systematic Reviews of Interventions [version 5.10 (updated March 2011)]. Retrieved from www.handbook.cochrane.org.

Thompson, A. D., Bell, L., Schrum, L., & Bull, G. (2008). Introduction. In A. D. Thompson, L. Bell, & L. Schrum (Eds.), *Framing research on technology and student learning in the content areas: Implications for educators* (pp. 1–12). Charlotte: Information Age Publishing.

*Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., … VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics. *Journal of Statistics Education*, *26*(2), 103–109. https://doi.org/10.1080/10691898.2018.1473061

Tishkovskaya, S., & Lancaster, G. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2), 1–56. https://doi.org/10.1080/10810730150501413

*Utts, J., Sommer, B., Acredolo, C., Maher, M. W., & Matthews, H. R. (2003). A study comparing traditional and hybrid internet-based instruction in introductory statistics classes. *Journal of Statistics Education*, *11*(3), 1–14.

Valente, J. C. (2009). *Handbook of research synthesis and meta-analysis* (2nd ed.; H. Cooper, L. V. Hedges, & J. C. Valente, Eds.). New York: Russel Sage Foundation.

Viechtbauer, W. (2010). Conducting meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112–125. https://doi.org/10.1002/jrsm.11.

Vo, H. M., Zhu, C., & Diep, N. A. (2017). The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies in Educational Evaluation*, *53*, 17–28. https://doi.org/10.1016/j.stueduc.2017.01.002

*Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology*, *92*(1), 137–143. https://doi.org/10.1037/0022-0663.92.1.137

*Wang, X. (1999). Effectiveness of statistical assignments in MPA education: An experiment. *Journal of Public Affairs Education*, *5*(3), 319–326. https://doi.org/10.1016/j.orggeochem.2011.11.003

Ward, B. (2004). The best of both worlds: A hybrid statistics course. *Journal of Statistics Education*, *12*(3). Retrieved from http://www.amstat.org/publications/jse/v12n3/ward.htm

Warren, S. J., Lee, J., & Najm, A. (2014). The impact of technology and theory on instructional design since 2000. In Michael J. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications technology* (4th ed., pp. 89–99). New York: Springer Science & Business Media.

Whiteman, N. (2003). *Research methods and computer-mediated- communication: A comparative study of a "real" and "virtual" learning environment*. University of London. Retrieved from http://www.pauldowling.me/studentswork/whiteman(2003).pdf

*Wilmoth, J., & Wybraniec, J. (2018). Profits and pitfalls: Thoughts on using a laptop computer and presentation software to teach introductory social statistics. *Teaching Sociology*, *26*(3), 166–178.