Journal of MultiDisciplinary Evaluation
Volume 11, Issue 24, 2015

ISSN 1556-8180
http://www.jmde.com

The Quality of Mathematics Education Technology Literature

Robert N. Ronau
University of Cincinnati

Christopher R. Rakes
University of Maryland, Baltimore County

Sarah B. Bush
Bellarmine University

Shannon O. Driskell
University of Dayton

Margaret L. Niess
Oregon State University

David Pugalee
University of North Carolina-Charlotte

Background: The present study evaluated the quality of 1,165 scholarly literature papers about mathematics education technology literature.

Purpose: The purpose of the present study was to determine the extent to which mathematics education technology literature reports the information needed to support the scientific basis of a study.

Setting: N/A

Intervention: N/A

Research Design: A systematic review was used to organize the data collection and analysis processes

Data Collection and Analysis: A literature search was conducted to identify scholarly papers that addressed the use of technology in mathematics education. A coding process was developed to record descriptive information about each paper. The Quality Framework developed for this process provided a structure to identify key information across research types based on types of analyses conducted, assigning a certain number of possible points based on the type of research conducted.

Findings: Dissertations accounted for a surprisingly high portion of the literature and research: 39.7% of the available literature and 57.0% of the research studies. The overall quality of the mathematics education technology literature was lower than we expected, averaging only 48.9% of the points possible. We noted that the quality of research papers, with respect to possible point values averaged 54.6% over four decades. For mathematics education technology researchers, manuscript reviewers, and editors, these results

suggest that more attention is needed on the information being included and excluded from scholarly papers, especially with regard to connections to theoretical frameworks and research designs.

Keywords: *education; mathematics education; technology.*

The present study evaluated the quality of mathematics education technology literature. The use of education technology in the teaching and learning of mathematics, a practice that has been studied, discussed, and promoted for decades as a fundamental principal of mathematics education (National Council of Teachers of Mathematics (NCTM), 1989, 2000), has resulted in unprecedentedly low computer to student ratios in schools nationwide (Peck, Cuban, & Kilpatrick, 2002). Peck et al. (2002) noted, however, that improved access is only half the battle: the other half is using the available technology in a way that improves student learning.

Koehler, Shin, and Mishra (2011) found that failure to address critical issues such as measure validity and reliability may impede teachers' efforts to use research findings to improve their use of technology in the classroom. Their finding echoed the position of the National Research Council: "The prevailing view is that findings from education research studies are of low quality and are endlessly contested—the result of which is that no consensus emerges about anything" (Shavelson & Towne, 2002, p. 28). A number of subsequent reports have indicated that education research has continued to be of inconsistent quality (e.g., Johnson & Daugherty, 2008; Tobin, 2007; Towne, Wise, & Winters, 2005), suggesting that the education technology research field needs more guidance to focus its efforts to improve quality, an undertaking which requires further research. Without high quality reporting of studies, the ability to engage in the scientific process of building knowledge from sets of studies is compromised (Shavelson & Towne, 2002); without research about quality, the prospects for improvement in the field are limited.

The present study identified characteristics necessary for mathematics education technology literature to be considered *high quality,* examined that literature for the presence of these characteristics, and identified potential leverage points for enhancing the quality of mathematics education technology literature. To clarify what we mean by *literature*, we refer to Shavelson & Towne's (2002) position that education research serves two purposes, to add to the field's understanding of education phenomena and to inform practical decision-making. In the mathematics education technology literature, research papers primarily serve the first purpose while non-research papers primarily (e.g., practitioner papers) serve the second. We included both research and non-research papers in our sample of *literature*, and we considered the inclusion of non-research papers to be a unique, important aspect of the study: As Shavelson and Towne (2002) pointed out,

> Another key problem has been the sharp divide between education research and scholarship and the practice of education in schools and other settings. This disconnect has several historic roots: researchers and practitioners have typically worked in different settings…teacher education has typically relied on practical experience rather than research. Operating in different worlds, researchers and practitioners did not develop the kinds of cross fertilization that are necessary in fields where research and practice should develop reciprocally…the expectation that research-based information will be available and should be part of the decision-making process needs to be cultivated both in the public and in the research community…." (Shavelson and Towne, 2002, pp. 14-15, 96)

We consider the expectation that non-research papers connect classroom practice to current research to be reasonable based on Shavelson and Towne's admonitions. Furthermore, such connections are critical if the research that is produced is to produce improvement in teaching practices. We therefore included non-research papers in our sample and developed quality criteria relevant to such papers.

Purpose and Research Questions

The purpose of the present study was to evaluate the quality of mathematics education technology literature. We emphasize that the purpose was not to examine the effectiveness of any particular technology. Although an examination of how

technology affects various education outcomes is an important topic of study, such investigations typically examine a particular type of technology and focus on how the technology was integrated. Our focus here is broader: Synthesizing the literature through an evaluation of quality.

A synthesis of literature quality provides a critical first step in providing a structure for considering how individual study results are moderated or mediated by the quality of the reporting. We began this synthesis by developing a framework to organize the attributes most commonly associated with high quality, which we refer to as the Quality Framework (QF). Next, we developed a quantitative measure of quality based on the categories of the QF. With this framework and measure developed, the present study addressed seven questions:

1. What types of research and papers are available in the body of mathematics education technology literature?
2. What is the quality of the mathematics education technology literature, as measured by the Quality Framework?
3. Does the quality of mathematics education technology literature differ across research types?
4. Does the quality of mathematics education technology literature differ across publication types?
5. Did the quality of mathematics education technology research literature change over time?
6. Did the quality of mathematics education technology non-research literature change over time?
7. If there are differences in quality across research and/or publication types, can these differences be traced to particular categories of the Quality Framework?

Question 1 served as a starting point for the present study to provide a broad view of the education technology literature landscape. Question 2 formed the foundation of the study, providing a unique QF score for each paper in our sample of literature. Questions 3-7 examined potential differences across important characteristics of literature.

Background Literature and Conceptual Framework

Scientific Principles of Research

Shavelson and Towne (2002) offered six scientific principles to guide the development of high quality research: (1) Pose significant questions that can be investigated empirically; (2) Link research to relevant theory; (3) Use methods that permit direct investigation of the question; (4) Provide a coherent and explicit chain of reasoning; (5) Replicate and generalize across studies; and, (6) Disclose research to encourage professional scrutiny and critique.

We used these principles as a foundation for considering the quality of mathematics education technology literature, building on an assumption that no single study can capture the breadth and depth needed to address complex issues in education. We also recognize, however, that higher quality studies (i.e., rigorous, well-designed and reported) offer more valid, reliable information to contribute than studies that fail to report important information. For example, Dynarski et al. (2007) responded to the Shavelson and Towne (2002) principles by investigating the efficacy of a number of technology products (e.g., PLATO Achieve, iLearn Math) to improve student achievement in mathematics. The products being investigated were used primarily to offer students tutorials and drill-and-practice computer activities. The purposive sample consisted of approximately 9,400 students in the classrooms of 428 teachers within 132 schools across 33 districts, and the treatment groups were randomly assigned to a specific technology or control group. Dynarski et al. used multiple measurement tools such as classroom observations, teacher surveys, interviews, and a standardized student achievement measure (SAT-10). They found that these technology programs showed no significant improvement in student test scores. Some readers have interpreted such results to mean that technology does not help in the learning of mathematics, but such a conclusion is premature and does not account for the complexity of education research that cannot be captured by any single methodology or by any individual study. Specifically, the Dynarski et al. study was not able to address questions about how the tutorials and drill-and-practice software were used by the teachers to improve student achievement, nor did it measure the degree of implementation fidelity with respect to teacher training, including how

pedagogy and classroom discourse impacted the role teachers played in the study (Fitzer et al., 2007). Factors such as these have been found to influence the efficacy of technology use in mathematics education for improving student achievement (Pape et al., 2011) and may have been influenced by the results from the Dynarski et al. study.

From such examples, we gain an understanding that the accumulation of evidence from research generated through multiple studies is far more powerful than the evidence from individual studies for considering whether and how technology should be used to enhance the teaching and learning of mathematics. In one such synthesis of studies, Burrill et al. (2002) examined 43 studies on calculator and graphing calculator use in the classroom to identify the most effective ways to integrate handheld technologies to improve student outcomes. They found that the research on handheld technologies was not of high quality, often missing key information such as the type of handheld technology used, the content being learned with the handheld technology, and the ways the technology was used (e.g., instruction, assessment). Ellington (2003, 2006) followed a model similar to the recommendations of Burrill et al. when she conducted meta-analyses of 54 calculator studies and 42 graphing calculator studies to determine the effect of calculator and graphing calculator integration on student achievement and attitudes toward mathematics. She found that both calculators and graphing calculators had the greatest positive effects when the technologies were integrated into both assessment and instruction. She also found that graphing calculators were moderately effective when integrated into instruction but not assessment.

Quality Framework

The Quality Framework (QF), originally developed by Ronau et al. (2010), used the scientific principles to organize key information needed to support the scientific basis of a study. Specific categories were determined by consulting literature on education research design to develop a structure to identify key information across a number of research types based on the types of analyses being conducted. For example, qualitative (Creswell, 2009; Patton, 2002), quantitative (Shadish, Cook, & Campbell, 2002), and mixed method analyses (Shadish et al., 2002; Teddlie & Tashakkori, 2009) were examples of key types of research methodologies addressed and the sources we used to support the reporting of critical information for them. The QF components were chosen to honor a wide array of purposes and methods. To minimize subjectivity, the accompanying measure focused on the presence or absence of relevant components and did not assess how well the components were addressed, nor did we rate the appropriateness of design and measurement choices. For example, we did not rate whether studies using instruments with multiple forms reported alternate-forms reliability, only whether reliability was addressed (1 point if addressed, 0 points if not addressed). This measure therefore serves as a critical first step toward establishing a reliable and valid method for discussing quality, approaching quality from the perspective that a missing component cannot contribute to a paper's quality. Such an approach provides at least two advantages for a first step. First, overall QF scores provide meaningful signposts for the field as a whole. Second, QF sub-scales and scores provide guidance for future researchers to improve quality by including key components that have not been typically reported.

QF was developed with the recognition that the purposes of different types of research and papers require different components: the quality should therefore be evaluated differently as well. So although many of the characteristics across research types overlap, the QF components included for each research type are distinct, yielding different point totals. The most useful comparison across research types for the QF measure is how well each addresses its relevant components, not on whether one type of methodology is better than another.

QF includes a number of characteristics such as the research design, threats to internal, external, construct, and statistical conclusion validity, trustworthiness, and group assignment and selection mechanisms (Ronau et al., 2010). The initial version of the framework was evaluated through two rounds of peer debriefing. Based on that feedback, the framework was revised to include whether studies explicitly provided a purpose statement and research questions or hypotheses. We also measured the extent to which the studies were grounded in extant literature and guided by a theoretical framework (Rakes, 2012).

QF is divided into three sets of components: Theoretical Connections, Measurement Trustworthiness, and Design Clarity and Validity. While research syntheses and meta-analyses have many features that set them apart from other methodologies, their analytic techniques follow quantitative, qualitative, or mixed method

procedures (Cooper, 1998; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001), so separate categories were not created for them within QF. Similarly, design experiments can also be quantitative, qualitative, or mixed methods, so separate categories were not created for them either. However, all design experiments in our sample presented only qualitative results, so design experiments were included only under qualitative research in the present study.

The inclusion of non-research papers (descriptions of classroom activities or strategies, anecdotal descriptions, book reviews, and opinion papers) is an unprecedented and important feature in QF. The ability of such literature sources to offer unique and useful connections between research and practice is a priority in education policy decisions (Easton, 2010). Connecting research to previous research is a fundamental part of the scientific method, as Shavelson and Town (2002) discussed in Scientific Principles 2, 5, and 6; but connecting research to decisions made in the classroom, in schools, across states is also critical. Papers published in practitioner journals should explicitly connect recommendations, activities, and assessments to the current body of research in order to inform stakeholders and decision makers outside the academy. For example, the *Mathematics Teacher*, a practitioner-oriented journal, describes its purposes as providing a "forum for sharing activities and pedagogical strategies, deepening understanding of mathematical ideas, and linking mathematics education research to practice" (National Council of Teachers of Mathematics, 2014, n.p.).

Education researchers have long attempted to make their research accessible to teachers. For example, dating back to the 1970s, Kennedy (1997) described the expansion of research methodologies to include approaches such as case studies and ethnographies in an effort to better reflect teachers' views of the classroom. On the flip side, she noted, "research that is conceptually accessible to teachers may be research that does *not* [sic] challenge assumptions to introduce new possibilities" (p. 10). She also stated that as research has made more connections to practice, researchers "have discovered the intransigence of prior beliefs, the frequent popularity of untested fads, and the frequent lack of receptivity to tested ideas" (p. 10).

When authors target a practitioner audience, they should therefore make concerted efforts to connect those articles to research and to provide clearly articulated actions to help practitioners integrate the research into their practice. On the flip side, when practitioners decide to use materials in their classroom, they should rely on research-based materials and practices rather than seeking the current most-popular fads and trends. Moreover, providing theoretical support for a proposed practice is important regardless of whether the author is a researcher or practitioner. The quality of the evidence presented to practitioners is a critical support for such a cycle. We therefore included an examination of non-research paper quality in the present study.

In earlier iterations of the present study, we found in a subset of our current sample (n = 309) that non-research papers constituted a large portion (44%) of the papers found in the literature (Ronau et al., 2010). Because the subsample was randomly chosen, we assumed that the ratio of non-research papers was consistent with the overall sample. With such a large portion of the literature being non-research, we concluded that any study of literature quality must include these papers.

Figure 1 presents our current conceptualization of QF, by research type: quantitative, qualitative, mixed methods, and other. Other includes non-research papers, theory development, literature reviews, and descriptions of technology development for mathematics education. The framework divides important components of reporting into three categories: Theoretical Connections, Design and Validity, and Validity and Reliability.

Theoretical Connections. The Theoretical Connections category addresses Shavelson and Towne's (2002) Scientific Principle 2 (*linking research to theory*) by measuring how well the study was grounded in the literature. It addresses Scientific Principle 4 (*Providing a coherent and explicit chain of reasoning)* by determining the degree to which the study was connected to a conceptual framework. Theoretical connections are fundamental for enabling scholarly papers to serve as foundations for future research and for informing practice (Congdon & Dunham, 1999). We applied a set of definitions to minimize the subjectivity of our rating. A paper was considered "well supported" (2 points) if it presented evidence for how the literature base guided the development of the purpose, "partially supported" (1 point) if it presented a literature base but did not explicitly connect it to the development of the purpose, or "not supported" (0 points) if it did not present relevant literature to support the purpose. A "well connected" paper (2 points) presented a theoretical framework and described how the framework guided the purpose/procedures and,

for research papers, the interpretation of results. A "partially connected" paper (1 point) presented a theoretical framework and either how it guided the development of the purpose/procedures or conclusions/recommendations, but not both. A "not connected" (0 points) paper either did not present a theoretical framework or presented one but did not describe how it guided the development of the purpose/procedures nor was it used to guide the development of conclusions or recommendations.

QF focuses on Theoretical Connections for non-research papers. For example, Abramovich & Ehrlich (2007) published a computer activity designed to address misconceptions in solving inequalities. This paper was a report of an instructional activity and was classified as non-research. The paper was; however, well grounded in the literature and well connected to a conceptual framework which scored all of the possible five points for non-research papers. de Villiers (2004) shared strategy for using dynamic geometry to enhance teachers' understanding (and thereby their instruction) of proof. This paper also scored all five points of the Theoretical Connections category by providing a strong grounding in the literature and a well-connected framework.

Design and Validity. The Design and Validity component addresses Shavelson and Towne's (2002) Scientific Principle 1 (*Pose significant questions that can be investigated empirically*) by determining if the paper contained an explicit purpose statement and research questions or hypotheses. One point was awarded for each of these present. Scientific Principle 3 (*Use methods that permit direct investigation of the question)* was addressed by coding reporting of the research design and threats to the validity of the study. Quantitative papers were scored for Design Robustness (up to 3 points, see Figure 1). Studies using a control group received one point. Studies conducting true (randomized) experiments or regression discontinuity designs each received 2 additional points. Quasi-experiments were also able to receive the full three points in this category: Random and purposive (probabilistic) sampling strategies were assigned 2 points, and if they included a control group, they scored at the same level as true experiments (3 points). We adopted this point assignment strategy to ensure that the quality measure reflects an acknowledgement that unique situations arise in education research that may prevent or make random assignment unfeasible. Convenience and unclear sampling strategies were assigned 1 point. For the category Threats to Validity Addressed, we coded whether a study addressed the four types of validity threats identified by Shadish et al. (2002): internal, external, construct, and statistical conclusion validity (1 point each).

As with quantitative studies, the Design Clarity and Validity component for qualitative studies measured whether a purpose statement and research questions or hypotheses were included. Internal, external, and construct validity threats were also considered relevant for qualitative studies (Creswell, 2007).

<table>
<tr><th colspan="2">Mixed Methods (up to 17 pts)</th><th rowspan="2">Theory Development Papers and Literature Reviews (up to 6 pts)</th></tr>
<tr><th>Quantitative (up to 15 pts)</th><th>Qualitative (up to 12 pts)</th></tr>
<tr>
<td><u>Theoretical Connections (up to 4 pts)</u>
• Literature Support (≤ 2 pts)
➢ Well Grounded (2 pts)
➢ Partially Grounded (1 pt)
➢ Not Grounded (0 pts)
• Conceptual Framework Connections (≤ 2 pts)
➢ Well Connected (2 pts)
➢ Partially Connected (1 pt)
➢ Not Connected (0 pts)
<u>Design and Validity (up to 9 pts)</u>
• Purpose Statement (1 pt)
• Research Questions/Hypotheses (1 pt)
• Design (up to 3 pts)
➢ Randomized Experiment (2 pts)
➢ Regression Discontinuity Design (2 pts)
➢ Quasi-Experimental Design with:
▪ Sampling Strategies Unclear (1 pt)
▪ Convenience Sample (1 pt)
▪ Other Sampling Strategies (2 pts)
➢ Use of Control Group (1 pt)
• Threats to Validity Addressed (up to 4 pts)
➢ Internal (1 pt) ➢External (1 pt)
➢ Construct (1 pt) ➢Statistical Conclusion (1 pt)
<u>Validity and Reliability (up to 2 pts)</u>
• Reliability (1 point)
➢ Internal Consistency ➢Split Half
➢ Inter-Rater ➢Test-Retest
➢ Alternate Forms
• Validity (1 point)
➢ Content ➢Construct
➢ Concurrent Criterion ➢Discriminant
➢ Predictive Criterion ➢Convergent</td>
<td><u>Theoretical Connections (up to 4 pts)</u>
• Literature Support (≤ 2 pts)
➢ Well Grounded (2 pts)
➢ Partially Grounded (1 pt)
➢ Not Grounded (0 pts)
• Conceptual Framework Connections (≤ 2 pts)
➢ Well Connected (2 pts)
➢ Partially Connected (1 pt)
➢ Not Connected (0 pts)
<u>Design and Validity (up to 5 pts)</u>
• Purpose Statement (1 pt)
• Research Questions/Hypotheses (1 pt)
• Design (1 pt)
➢ Biography
➢ Phenomenology
➢ Historical/Narrative
➢ Grounded Theory
➢ Ethnography
➢ Case Study
• Threats to Validity Addressed (up to 3 pts)
➢ Internal (1 pt) ➢External (1 pt)
➢ Construct (1 pt)
<u>Validity and Reliability (up to 2 pts)</u>
• Reliability (1 point)
➢ Internal Consistency ➢Inter-Rater
• Validity (1 pt)
➢ Persistent Observation ➢ Member Checks
➢ Triangulation ➢ Thick Description
➢ Peer Debriefing ➢ Dependability Audit
➢ Negative Case Analysis ➢ Confirmability Audit
➢ Referential Adequacy ➢ Reflective Journal</td>
<td><u>Theoretical Connections (up to 4 pts)</u>
• Literature Support (up to 2 pts)
➢ Well Grounded (2 pts)
➢ Partially Grounded (1 pt)
➢ Not Grounded (0 pts)
• Conceptual Framework (up to 2 pts)
➢ Well Connected (2 pts)
➢ Partially Connected (1 pt)
➢ Not Connected (0 pts)
<u>Design and Validity (up to 2 pts)</u>
• Purpose Statement (1 pt)
• Design (1 pt)
➢ Biography ➢Phenomenology
➢ Grounded Theory ➢Ethnography
➢ Historical/Narrative ➢ Case Study</td>
</tr>
<tr><td colspan="2"></td><td>Other (up to 5 pts)</td></tr>
<tr><td colspan="2"></td><td><u>Theoretical Connections (up to 4 pts)</u>
• Literature Support (up to 2 pts)
➢ Well Grounded (2 pts)
➢ Partially Grounded (1 pt)
➢ Not Grounded (0 pts)
• Conceptual Framework (up to 2 pts)
➢ Well Connected (2 pts)
➢ Partially Connected (1 pt)
➢ Not Connected (0 pts)
<u>Design and Validity (up to 1 pt)</u>
• Purpose Statement (1 pt)</td></tr>
</table>

Figure 1. Categories included in the Quality Framework (QF) and points available within each category. Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other includes non-research papers and descriptions of mathematics education technology development.

Validity and Reliability. The Validity and Reliability component also addressed Shavelson and Towne's (2002) Scientific Principle 3 by measuring whether a study addressed the reliability and validity of measures (Urbina, 2004). A point was assigned for addressing one or more of five types of reliability: internal consistency, split-half, test-retest, inter-rater and alternate form of reliability. An additional point was assigned for reporting one or more of six types of validity: content, concurrent criterion, predictive criterion, construct, discriminant, and convergent.

We found that internal consistency and inter-rater reliability were also addressed by a small number of qualitative studies (n = 7), indicating that reliability was an appropriate measure to include for qualitative studies. The validity of the study (1 point; i.e., *credibility* in Creswell, 2007; *trustworthiness* in Patton, 2002; Lincoln & Guba, 1985) was coded by examining ten categories: persistent observation, triangulation, peer debriefing, negative case analysis, referential adequacy, member checks, thick description, dependability audit, conformability audit, and reflective journal.

Mixed Methods Studies in the Quality Framework

Mixed methods studies included all of the components unique to both quantitative and qualitative studies. Literature Support, Framework/Theory Connections, Purpose Statement, Research Questions/Hypotheses, Threats to Validity, and Reliability were coded for the overall study (that is, not separately for the quantitative and qualitative components). Due to the overlap of the categories for qualitative and quantitative studies, mixed methods studies could earn up to 17 points: the 15 points available to quantitative studies, an additional point for addressing validity of the qualitative measure and another for addressing the qualitative design explicitly and distinct from the quantitative design.

Method

A literature search was conducted to identify scholarly papers that addressed the use of technology in mathematics education. A coding process was developed to record descriptive information about each paper and to minimize subjectivity across coders. This process began with a literature search strategy, development of a coding tool, validating the coding tool, and training coders to use the coding tool reliably.

Literature Search Strategy

To conduct our literature search, we adopted the strategies outlined by Lipsey and Wilson (2001) for making a search systematic (e.g., a priori determination of inclusion criteria, comprehensive database searching, searching of bibliographies for non-indexed papers, and inclusion of "gray" literature such as dissertations). We required papers to be relevant to three criteria to be retained in the sample. The paper must be about (1) technology (using search terms such as *technology*, *calculators*, *computers*), (2) mathematics (e.g., *mathematics*, *algebra*, *geometry*, *visualization*, *representation*), and (3) education (e.g., *education*, *teaching*, *learning*). Although we did not limit the sample to papers published in the U.S. or papers about education in the U.S., we did exclude papers if we were unable to retrieve the full text or if the papers were unavailable in the English language, only because we were unable to code them. When papers were indexed in English in a database but the full text was in a different language (n = 11), we contacted the authors to try to obtain an English copy. Because of this effort, three papers that would have been excluded were able to be retained in the sample.

The electronic platforms searched were EBSCOWeb (databases included ERIC, Academic Search Premier, PsychInfo, Primary Search Plus, Middle Search Plus, Educational Administration Abstracts), JSTOR (limited to the following disciplines: Education, Mathematics, Psychology, and Statistics), OVID, ProQuest (Research Library, Dissertations & Theses, Career & Technical Education), H. W. Wilson Web (Education Full Text), and Google Scholar. In addition to the electronic searches, the bibliographies of relevant papers were searched to identify other potentially relevant papers, yielding an additional 98 papers. Although we recognize that other relevant papers may exist, Lipsey and Wilson (2001) posited that a systematic, comprehensive literature search is the best defense for minimizing publication bias and maximizing the representativeness of the sample. The inclusion of gray literature, research that is accepted based on its scientific merits rather than significance of findings (Rothstein & Hopewell, 2009), is a well-supported defense against publication bias (e.g., Song, Hooper &Yoon, 2013; Conn, Valentine, Cooper & Rantz, 2003). We

therefore included gray literature such as dissertations, master's theses, and technical reports. Our literature search identified 1,427 papers potentially relevant to mathematics education technology, from 1968 to 2009.

A total of 215 papers were removed from the 1,427 because, although the titles indicted that they were potentially relevant, closer inspection revealed that the studies did not address technology, or mathematics, or education. Additionally, conference papers were not always indexed in the electronic databases as individual papers; instead, they were listed only as part of conference proceedings, which meant that locating the relevant papers required culling the table of contents of each proceedings document. Although we did do so for every proceeding that was identified by the electronic searches, we recognized that many conference proceedings do not index their papers electronically, nor did we have a systematic way to determine which conference papers were subsequently published (i.e., titles might be slightly different although reporting on the same study, author lists might change between the conference presentation and final publication). Additionally, although journal papers and dissertations have been systematically indexed over the decades, either via microfiche, card catalogs, or monographs, no such indexing systematically occurred for conference proceedings. We therefore recognized that the inclusion of conference papers would possibly bias our sample and possibly reduce its representativeness of the population. We therefore decided to exclude the 47 conference papers that we did find.

This literature search strategy resulted in a final sample of 1,165 papers, consisting of journal articles, full books, book chapters, technical reports, master's theses, and doctoral dissertations. Dissertations were the most common source of gray literature in the sample (n = 480), and master's theses were the second most common (n = 75). If one or more journal articles reported results from a dissertation study or master's thesis, only the journal articles were included.

Coding Tool Development

We created a Microsoft Access database to organize the coding process. This database allowed us to create dropdown lists and checkboxes for indicator variables to minimize variation of responses for similar information. It also provided a mechanism to track the stage of coding for each paper (retrieved, waiting on Inter-Library Loan, coded, feedback ready, complete), which helped ensure that papers were not missed or only partially coded. The 246 fields in the database consisted of categories for descriptive information about the paper (bibliographic information, content area, grade level, type of publication, theoretical connections, purpose of paper, type of research), a place to list of any research questions investigated in the paper as well as theoretical frameworks and their sources, data sources used (performance assessments, surveys, journals, observations, interviews, focus groups, and content analyses), outcomes studied (student achievement, learning, orientation, behavior teacher knowledge, orientation, behavior analysis of an instrument report of classroom or teaching activity research to practice), types of technology (calculators, probeware, computer software, Internet resources), connections to the NCTM (2000) principles (Learning, Teaching, Equity, Technology, Curriculum, Assessment), connections to Technological Pedagogical Content Knowledge (Mishra & Koehler, 2006; Niess, 2005), research design, reliability, validity, number and types of instrumentation, sample information, and threats to validity. These indicator variables were chosen to provide a broad view of the information available from mathematics education technology literature.

Coding Tool Validation and Refinement

The coding tool was refined through an iterative process and was first piloted with three papers and two coders. Refinements based on the results of this pilot test were examined with all six researchers coding the same, original three papers. This process was repeated through three more iterations of refinement and the coding of 27 more papers (i.e., 30 articles were coded by the whole group). Publication, research, technology, data source, and outcome types were identified through content analysis informed by a grounded theory approach. The categories of each type evolved as coding progressed until a saturation point was reached (i.e., few additional categories emerged, grouped as "other").

At each stage of the refinement process, definitions and criteria for making coding decisions were clarified through weekly training sessions, but we quickly realized that the application of those definitions and criteria to specific papers was problematic. We also realized that many of the decisions being made during the coding process were subject to interpretation. We

concluded that a random sample of double-coded papers would be insufficient for ensuring inter-rater agreement and adjusted our coding strategy to require that all papers be coded by at least two coders and that all coders worked with every other coder.

We used a counterbalanced design to assign coders to papers, minimizing potential coder bias. First, each coder was paired with each of the other coders (i.e., six coders, each paired with the other five for a total of 15 coding teams). A random number was computed for each paper in Microsoft Excel, papers were then listed in a random order, and teams were assigned to papers systematically. Within each team, each member was designated as the primary coder for a set of papers and as the secondary coder for their partners' studies. Once primary coders completed coding a paper, the secondary coder was notified, who then reviewed, confirmed, or questioned each of the coding decisions.

Coder Training and Alignment

Coders were trained to recognize and interpret the components of the papers correctly by whole group reviews of individual coding. For example, regular team discussions led to a number of coding clarifications to improve team alignment. Clarifications were needed to increase reliability on the coding of meta-analyses, mixed methodology, single subject research, action research, and survey research designs. Sampling issues such as how to differentiate between purposive and convenience sampling were also discussed in the regular team meetings. How to code various types of measurement tools, instruments, and forms of data was discussed. Agreed-upon definitions and procedures were recorded and posted on the team website. Cross-validation by the second coder ensured adherence to the team decisions and consistent interpretation of specific papers. Any disagreements between the primary coder and secondary coder were automatically recorded in the database as part of the coding process and were subsequently discussed by the full team. Of the original 1,427 potentially relevant papers, there was full initial agreement (i.e., no disagreements) on 390 of them (approximately 27% agreement rate). This full agreement count is based on the total number of potentially relevant papers rather than the final sample size (1,165 papers) because decisions about relevance were included in the coding process. Of the 1,037 papers that required some level of discussion before finalizing the coding, the number of discussion points and clarifying questions raised by the second coder ranged from 1 to 12, Mean = 2.9, Median = 3 per paper (see Figure 2). Given that 246 fields were coded for each paper, we considered our consistency between coders within teams to be high (96.1% to 99.6% initial agreement). Even with strong consistency between coders for each paper, the high number of papers requiring some level of discussion supports our decision to double code every paper in the sample.

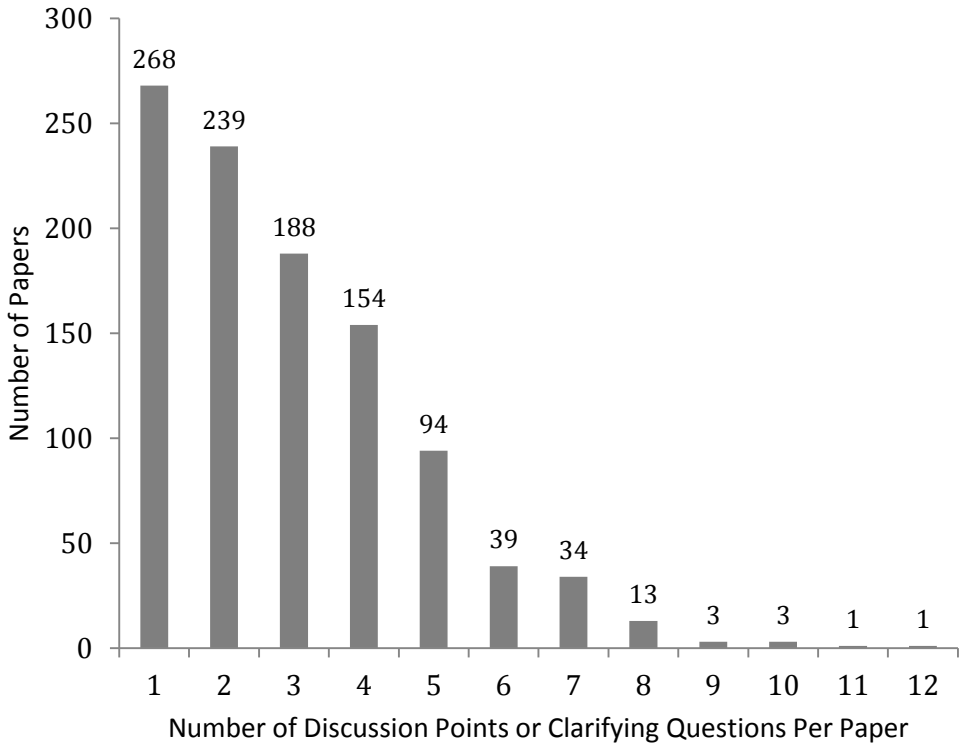

Figure 2. Histogram of the number of discussion points or clarifying questions per paper. N = 1,037 papers for which discussion or clarification were needed.

The 15 coding teams were highly similar in the amount of discussion and clarification needed to complete the coding of a paper. The average number of discussion points and clarification questions by team ranged from 2.0 to 4.5 per paper (98.2% to 99.2% initial agreement), Mean = 2.7, Median = 2.6. These low average amounts of discussion by paper and coding teams demonstrate a large degree of consistency in the application of definitions and criteria to the sample.

Types of Research and Papers Available

The number of papers in the 1960s, 1970s, and 1980s were low, with n = 2, n = 22, and n = 48 respectively. Nevertheless, we retained these papers in the sample because we believed that they may provide important insight for understanding the beginnings of mathematics education technology research.

Table 1
Number of Papers by Publication and Research Types Across Decades

Publication by Decade	Research Type					Publication by Decade Total
	Quantitative[a]	Qualitative[b]	Mixed Methods	Other Research[c]	Non-Research	
1960s (1960-1969)	**1**	**0**	**0**	**0**	**1**	**2**
Journals	0	0	0	0	1	**1**
Dissertations	1	0	0	0	0	**1**
1970s (1970-1979)	**10**	**1**	**1**	**0**	**10**	**22**
Journals	1	0	0	0	9	**10**
Dissertations	8	0	0	0	0	**8**
Other Publications[d]	1	1	1	0	1	**4**
1980s (1980-1989)	**25**	**4**	**2**	**2**	**13**	**46**
Journals	2	0	0	1	12	**15**
Dissertations	16	4	1	0	0	**21**
Other Publications[d]	7	0	1	1	1	**10**
1990s (1990-1999)	**113**	**45**	**58**	**12**	**75**	**303**
Journals	22	5	3	2	66	**98**
Dissertations	73	37	52	7	0	**169**
Other Publications[d]	18	3	3	3	9	**36**
2000s (2000-2009)	**218**	**154**	**108**	**51**	**261**	**792**
Journals	62	62	22	20	227	**393**
Dissertations	120	73	77	10	1	**281**
Other Publications[d]	36	19	9	21	33	**118**
Research Type Total	**367**	**204**	**169**	**65**	**360**	**1165**
Journals	**87**	**67**	**25**	**23**	**315**	**517**
Dissertations	**218**	**114**	**130**	**17**	**1**	**480**
Other Publications[d]	**62**	**23**	**14**	**25**	**44**	**168**

Note. Publication, research, and decade subtotals and totals are in boldface.
[a]Quantitative includes single subject designs and meta-analyses.
[b]Qualitative includes action research and design experiments.
[c]Other Research consists of theory development papers, literature reviews, and development of technology for mathematics education.
[d]Other Publications includes book chapters, full books, reports, and master's theses.

Non-research papers accounted for 360 of the 1,165 papers (30.1%), which left 842 research studies in the sample. The number of non-research papers was nearly equal to quantitative studies (n = 367) and was greater than both qualitative (n = 204) and mixed methods (n = 169) studies. Of the 204 qualitative studies, 123 (60%) were case studies. Quantitative studies included 195 surveys (53%), and mixed methods papers included 103 surveys (61%). Surveys and case studies were not often found in the Other Research category or in Non-research literature (see Figure 2).

Journal articles made up approximately 44.4% of the sample (Table 1), but 315 of them (60.9% of the 517 journal articles) were non-research papers, leaving the 202 research studies published in journals or a total of 25.1% of the 805 research studies. The 480 dissertations made up 41.2% of the sample (n = 1165), which accounted for 59.5% of the 805 research studies available in the mathematics education technology literature.

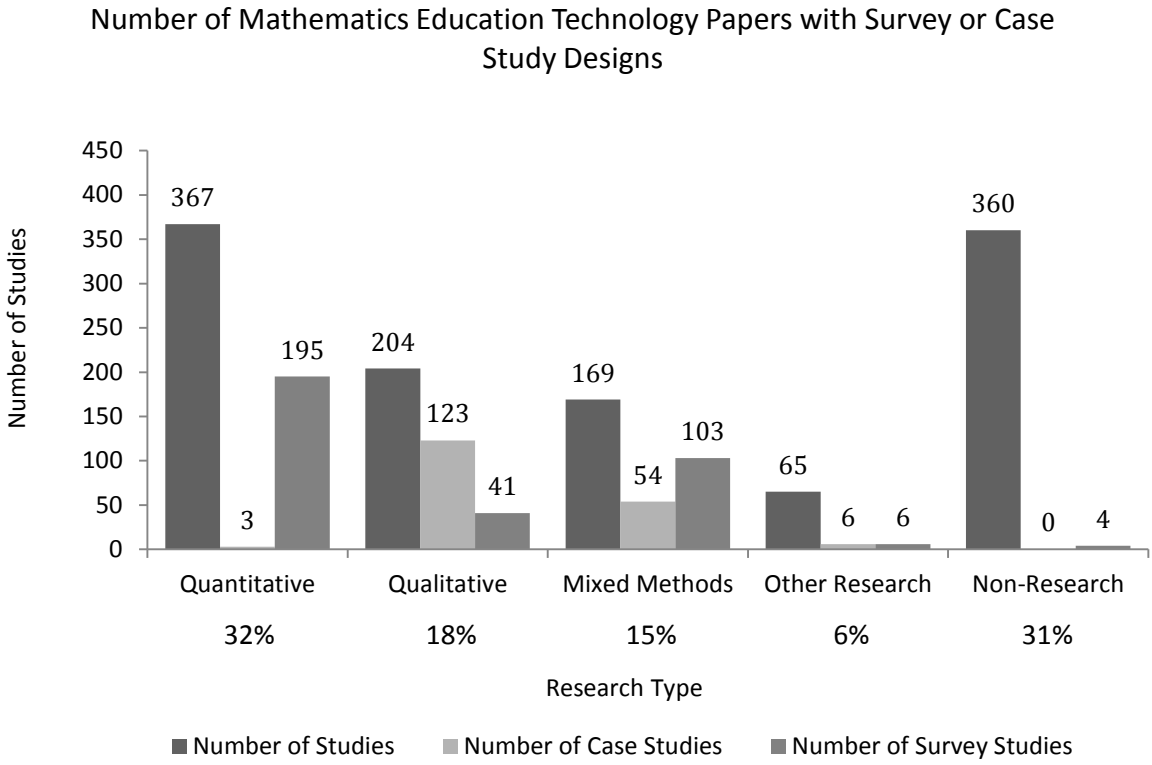

Figure 3. Number and percentage of papers that reported on survey data, by research type. Percentages are out of total number of papers by research type (see Table 1). Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other research includes theory development, literature reviews, and descriptions of technology development for mathematics education.

Quality Points Earned

Quality measures were computed from the QF (see Figure 1) as percentages of the number of points possible for the relevant design: five points for non-research papers, 12 for qualitative designs, 15 for quantitative designs, and 17 for mixed methods.

Overall, the sample papers addressed an average of 50.4% of the relevant QF components (Table 2). Other Research had the highest average quality percentage (86.4%) while non-research papers had the lowest (33.4%).

Table 2
Mean Percentage of Quality Points Earned by Publication and Research Types Across Decades

Publication Type by Decade	Research Type					Decade Mean
	Quantitative[a]	Qualitative[b]	Mixed Methods	Other Research[c]	Non-Research	
1960s (1960-1969)	**50.0**	**0.0**	**0.0**	**0.0**	**20.0**	**35.0**
Journals	0.0	0.0	0.0	0.0	20.0	20.0
Dissertations	50.0	0.0	0.0	0.0	0.0	50.0
1970s (1970-1979)	**55.6**	**18.2**	**50.0**	**0.0**	**26.0**	**40.2**
Journals	31.3	0.0	0.0	0.0	26.7	27.1
Dissertations	60.2	0.0	0.0	0.0	0.0	60.2
Other Publications[d]	43.8	18.2	50.0	0.0	20.0	33.0
1980s (1980-1989)	**54.3**	**45.5**	**46.9**	**55.7**	**32.9**	**47.0**
Journals	40.6	0.0	0.0	75.0	31.7	35.8
Dissertations	59.0	45.5	56.3	0.0	0.0	56.3
Other Publications[d]	48.3	0.0	37.5	36.4	40.0	45.0
1990s (1990-1999)	**55.3**	**51.3**	**60.1**	**52.4**	**32.8**	**49.8**
Journals	45.5	38.2	50.0	31.8	32.7	36.4
Dissertations	61.0	53.1	62.4	66.2	0.0	59.9
Other Publications[d]	46.9	50.9	42.5	38.6	33.3	42.4
2000s (2000-2009)	**54.1**	**50.6**	**59.4**	**62.1**	**35.3**	**48.4**
Journals	47.7	43.6	44.9	60.8	33.7	39.4
Dissertations	61.3	60.8	64.7	65.9	9.1	62.1
Other Publicationsd	43.0	39.6	51.7	61.6	44.8	46.3
Research Type Mean	**54.5**	**50.5**	**59.5**	**60.0**	**34.4**	**48.6**
Journals	**60.9**	**57.8**	**63.7**	**66.0**	**9.1**	**61.0**
Dissertations	**46.8**	**43.2**	**45.5**	**58.9**	**33.1**	**38.5**
Other Publications[d]	**44.8**	**40.6**	**48.3**	**57.2**	**41.3**	**45.0**

Note. — indicates that no papers of a particular publication and research type were published in a particular decade (see Table 1). Publication, research, and decade mean percentages are in boldface. *N* = 1,165apers.
[a]Quantitative includes single subject designs and meta-analyses.
[b]Qualitative includes action research and design experiments.
[c]Other Research consists of theory development papers, literature reviews, and development of technology for mathematics education.
[d]Other Publications includes book chapters, full books, reports, and master's theses.

Moderating Effects on Quality Points Earned

An analysis of variance revealed statistically significant differences in the percentage of relevant quality components addressed across research types, $F(4, 1205) = 133.9$, $p < .001$. Pairwise comparisons identified only one non-significant difference, which was between quantitative and mixed methods research, $p > .5$. All other pairwise comparisons were statistically significant, $p \leq .001$. From these analyses, we concluded that the degree to which Other Research (e.g., theory development, literature reviews) provided critical information for supporting future research and classroom practice was greater than

quantitative or mixed methods studies, which were in turn greater than qualitative studies. Such a conclusion does not mean that one type of research is better than another, only how well each type of research addresses the components necessary to convey important information about the study. Non-research papers had the lowest quality scores, indicating that the information provided in such papers about student, teacher, and other outcomes (e.g., instructional strategy descriptions) offer little support for helping teachers connect to research.

Research quality over time. Quality comparisons for research papers are shown in Figure 3 where the percent quality of publication types is compared across the decades. Dissertations show a stable average percent quality level hovering around 60% while journal research papers fluctuated around 45% after the 1970s. The papers in the "Other" category (i.e., book chapters, full books, reports, and master's theses) maintained an average quality level equivalent to that of research journal articles.

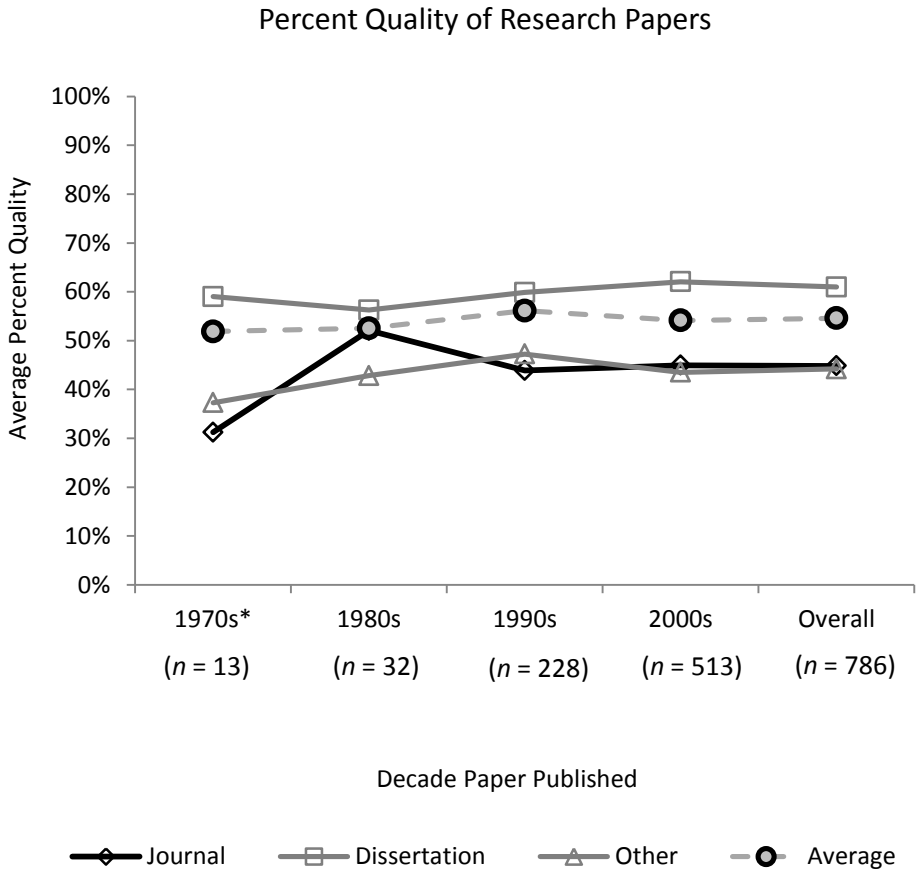

Figure 4. Average percent quality by decade and publication type. Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other Research includes theory development, literature reviews, and descriptions of technology development for mathematics education. Other Publications includes book chapters, full books, reports, and master's theses.
Note*: 1970* contains the two papers from 1968.

Non-research quality over time. Figure 4 illustrates the percent quality of non-research papers over four decades. Dissertations are not included in this figure because no dissertations were in the non-research category. The figure shows that journal non-research articles tended to

make stronger connections to the literature than papers in the "Other" category.

Together Figures 3 and 4 show that, on average, despite having to meet a greater set of indicators, research papers demonstrated a higher level of quality (54%) than non-research papers (37%). A notable feature of this set of papers is that dissertation studies remained at a fairly stable quality level hovering around 60% while journal research papers fluctuated around 50% after the 1970s. Not surprisingly, there were no dissertations in the non-research type of paper.

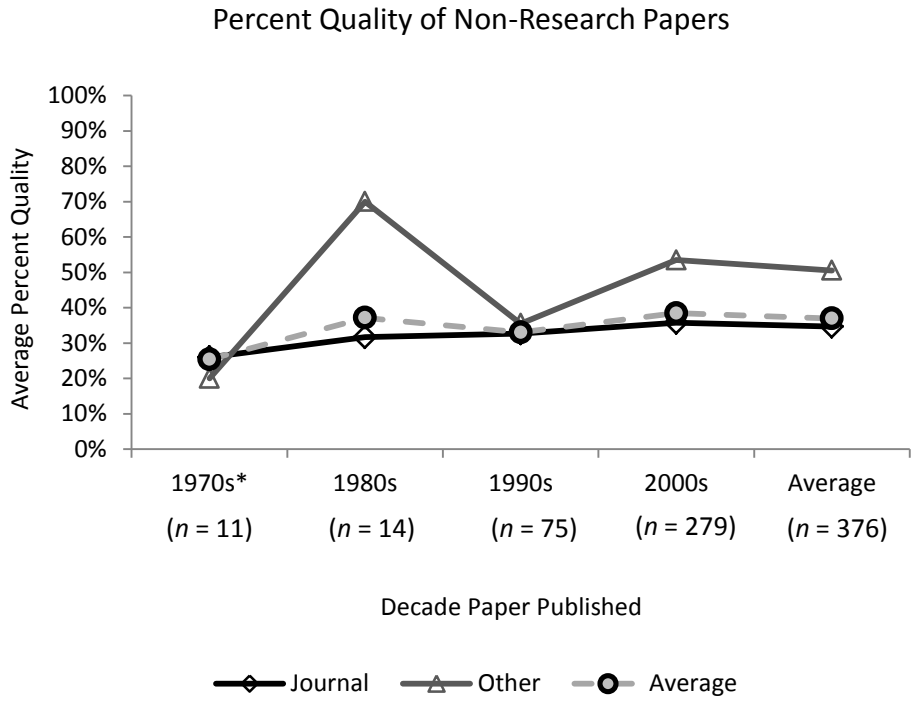

Figure 5. Average percent quality by decade and publication type. Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other Research includes theory development, literature reviews, and descriptions of technology development for mathematics education. Other Publications includes book chapters, full books, reports, and master's theses.

Quality Components

To understand why such differences might exist, we examined how each type of paper addressed theoretical connections and purpose statements, which were the QF components common across research types. Only 19 out of 1,165 papers (1.6%) did not provide an explicit purpose statement, so the analyses focused on Theoretical Connections, the degree of literature support and conceptual framework connections, which were each assessed up to two points (see Figure 1). Because there were two dependent variables to be analyzed across groups, and they were moderately correlated (r = .56, p < .001), a multivariate analysis of variance (MANOVA) was used to determine whether apparent differences across surveys, case studies, research types, and publication types were statistically significant. Such an approach minimizes risk of Type I error (Stevens, 2001). Hotelling's Trace, Pillai's Trace, and Wilks' Lambda were examined and found to provide consistent results. For the sake of space, only Wilks' Lambda is provided in Table 3.

Because case studies, surveys, publication types, and research types appeared to be the source of most patterns we observed in the data, we used these four variables as factors for the MANOVA on literature support and conceptual framework connections. The 15 sources of variance presented in Table 3 are the four main effects, six two-way interactions, four three-way interactions,

and one four-way interaction. Of the 15 potential effects, only three indicated statistical significance: the main effects of research type (quantitative, qualitative, mixed methods, other research, or non-research) and being a case study and the two-way interaction effect of publication with research type (journal, dissertation, or other publication).

Table 3
Results for MANOVA on Literature Support and Framework Connections Across Research Type, Publication Type, Case Studies, and Surveys

Source of Variance	Wilks' Lambda	F(DF)
[Case Studies]	**0.994**	**3.783 (2, 1,169)***
[Publication Type]	0.997	0.778 (4, 2,338)
[Publication Type] * [Case Studies]	0.996	1.061 (4, 2,338)
[Publication Type] * [Research Type]	**0.969**	**2.655 (14, 2,338)*****
[Publication Type] * [Research Type] * [Case Studies]	0.992	1.55 (6, 2,338)
[Publication Type] * [Research Type] * [Surveys]	0.990	1.163 (10, 2,338)
[Publication Type] * [Research Type] * [Surveys] * [Case Studies]	0.999	0.223 (4, 2,338)
[Publication Type] * [Surveys] * [Case Studies]	0.997	0.751 (4, 2,338)
[Publication Type] *[Surveys]	0.999	0.289 (4, 2,338)
[Research Type]	**0.981**	**2.888 (8, 2,338)****
[Research Type] * [Case Studies]	0.998	0.531 (4, 2,338)
[Research Type] * [Surveys]	0.995	0.665 (8, 2,338)
[Research Type] * [Surveys] * [Case Studies]	0.999	0.428 (2, 1,169)
[Surveys]	0.998	1.247 (2, 1,169)
[Surveys] * [Case Studies]	0.999	0.407 (2, 1,169)

Note. N = 1,210 papers. Names of variance sources are bracketed. Statistically significant effects are in boldface.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Literature Support. Tukey posthoc pairwise comparisons were computed for literature support points by each category of publication by research types as well as the main effects of research type and case study. For publication by research type comparisons, there were 14 categories, but one category (dissertations using other research) only had one paper, so it was eliminated from the pairwise comparisons. Therefore, each of the remaining 13 categories was paired with the other 12 for a total of 156 comparisons. Of the 156 comparisons, 86 (55.1%) were found to be statistically significant at the .05 alpha level. For research type comparisons, the five categories were paired with each of the other four for a total of 20 comparisons while for case study comparisons, there was only one pairwise comparison. Of the 20 comparisons for research types, 12 (60%) were statistically significant. Table 4 provides the average literature support points for each research, case study, and publication by research type category.

Table 4

Means and Standard Errors of Literature Support for Research Type Main Effects, Case Study Main Effects, and Publication by Research Type Interactions.

				95% Confidence Interval	
Description	N	M	SE	Lower Bound	Upper Bound
Research Types					
Qualitative	229	1.65	0.043	1.571	1.741
Mixed Methods	174	1.47	0.070	1.330	1.601
Other Research	47	1.16	0.199	0.770	1.551
Quantitative	392	0.74	0.042	0.661	0.831
Non-Research	368	0.58	0.041	0.500	0.660
Case Study Types					
Case Studies	193	1.66	0.087	1.49	1.83
Non-Case Studies	1017	1.27	0.048	1.18	1.37
Publication by Research Types					
Dissertation-Qualitative[a]	121	1.92	0.052	1.82	2.02
Dissertation-Mixed Methods	131	1.89	0.050	1.79	1.98
Dissertation-Quantitative[b]	227	1.77	0.038	1.70	1.85
Other Publication[c]-Other Research[d]	26	1.77	0.112	1.55	1.99
Journal-Other Research[d]	20	1.70	0.128	1.45	1.95
Journal-Qualitative[a]	72	1.58	0.067	1.45	1.72
Other Publication[c]-Qualitative[a]	36	1.47	0.095	1.29	1.66
Other Publication[c]-Mixed Methods	18	1.39	0.135	1.12	1.66
Journal-Quantitative[b]	91	1.25	0.060	1.14	1.37
Other Publication[c]-Quantitative[b]	74	1.23	0.067	1.10	1.36
Journal-Mixed Methods	25	1.08	0.114	.86	1.30
Other Publication[c]-Non-Research	59	0.70	0.074	.55	0.84
Journal-Non-Research	309	0.46	0.033	.40	0.52
Dissertation-Other Research[d]	1	0.00	—[e]	—[e]	—[e]

Note. Literature support maximum was 2 points.

[a]Qualitative includes action research and design experiments.

[b]Quantitative includes single subject designs and meta-analyses.

[c]Other Publications includes book chapters, full books, reports, and master's theses.

[d]Other Research consists of theory development papers, literature reviews, and development of technology for mathematics education.

[e]— indicates that the sample size was insufficient to compute the standard error and confidence interval.

Dissertations using qualitative methods scored significantly higher for literature support than dissertations using mixed methods or quantitative research and journals or other publications (e.g., book chapters, using other types of research (e.g., theory development, literature reviews), $p > .5$ (Figure 5). Journals using non-research, the factor with the lowest mean, were significantly lower than every publication by research type ($p < .001$) other than other publications using non-research ($p = .162$). Other publications using non-research were not significantly lower than journal articles using mixed methods ($p = .194$) but were significantly lower than all other factors ($p < .001$). Journal articles using mixed methods were not significantly different from journals using qualitative methods ($p = .291$) but were significantly lower than journals using other research methods ($p = .019$).

The degree of literature support was consistently lower for non-research papers ($p < .01$) and for dissertations was consistently higher ($p < .001$). Case studies provided literature support significantly more than non-case studies ($p < .001$).

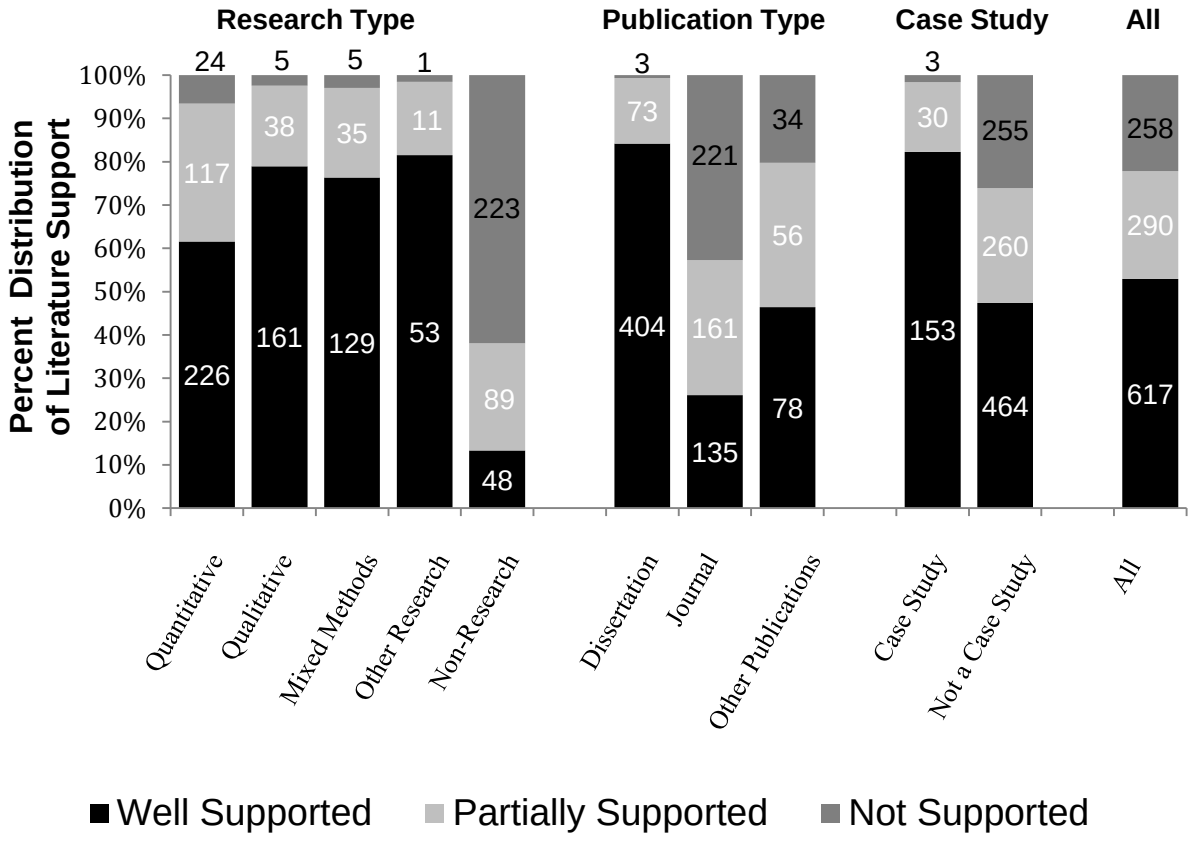

Figure 6. Percentage distribution of papers providing each degree of literature support by case study, research and publication type. Numbers within the stacked bars are the number of papers in each subset. *N* = 1,165 papers. Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other Research includes theory development, literature reviews, and descriptions of technology development for mathematics education. Other Publications includes book chapters, full books, reports, and master's theses.

Conceptual Framework Connections. Tukey posthoc pairwise comparisons were also computed for conceptual framework connection points by each category of publication by research types as well as the main effects of research type and case study. Of the 156 comparisons for publication by research type, 84 (53.8%) were found to be statistically significant at the .05 alpha level. Of the 20 comparisons for research types, 16 (80%) were statistically significant. Table 5 provides the average conceptual framework connection points for each research, case study, and publication by research type category.

As with literature support, conceptual framework connections were the lowest for non-research journal articles, and all but one pairwise comparison was statistically significant ($p < .05$) (Figure 6). Non-research journals were not significantly lower than non-research in other publications ($p = .084$). Dissertations using qualitative and mixed methods research were not significantly different ($p = .110$), but both scored significantly higher than dissertations using quantitative methods ($p < .05$).

Conceptual framework connections were consistently lower for non-research papers ($p \leq .001$) and for dissertations was consistently higher ($p < .001$). Case studies provided conceptual framework connections significantly better than non-case studies ($p < .001$).

Table 5

Means and Standard Errors of Conceptual Framework Connections for Research Type Main Effects, Case Study Main Effects, and Publication by Research Type Interactions.

				95% Confidence Interval	
Description	N	M	SE	Lower Bound	Upper Bound
Research Types					
Qualitative	229	1.26	0.055	1.15	1.36
Other Research	47	0.99	0.108	0.78	1.20
Mixed Methods	174	0.96	0.089	0.79	1.14
Quantitative	392	0.78	0.147	0.50	1.07
Non-Research	368	0.27	0.051	0.17	0.37
Case Study Types					
Case Studies	193	1.21	0.111	0.99	1.42
Non-Case Studies	1017	0.78	0.027	1.32	1.42
Publication by Research Types					
Dissertation-Qualitative[a]	121	1.45	0.068	1.32	1.59
Dissertation-Mixed Methods	131	1.30	0.067	1.17	1.43
Other Publication[c]-Other Research[d]	26	1.23	0.142	0.95	1.51
Journal-Qualitative[a]	72	1.16	0.087	0.99	1.33
Other Publication[c]-Qualitative[a]	36	1.15	0.122	0.91	1.39
Journal-Quantitative[b]	91	1.07	0.259	0.56	1.58
Other Publication[c]-Mixed Methods	18	0.91	0.205	0.51	1.31
Journal-Other Research[d]	20	0.75	0.162	0.43	1.07
Dissertation-Quantitative[b]	227	0.70	0.257	0.20	1.21
Journal-Mixed Methods	25	0.67	0.155	0.37	0.98
Other Publication[c]-Quantitative[b]	74	0.38	0.084	0.21	0.54
Other Publication[c]-Non-Research	59	0.36	0.094	0.17	0.54
Journal-Non-Research	309	0.18	0.041	0.1	0.26
Dissertation-Other Research[d]	1	0.00	—[e]	—[e]	—[e]

Note. Conceptual Framework Connections maximum was 2 points.

[a]Qualitative includes action research and design experiments.

[b]Quantitative includes single subject designs and meta-analyses.

[c]Other Publications includes book chapters, full books, reports, and master's theses.

[d]Other Research consists of theory development papers, literature reviews, and development of technology for mathematics education.

[e]— indicates that the sample size was insufficient to compute the standard error and confidence interval.

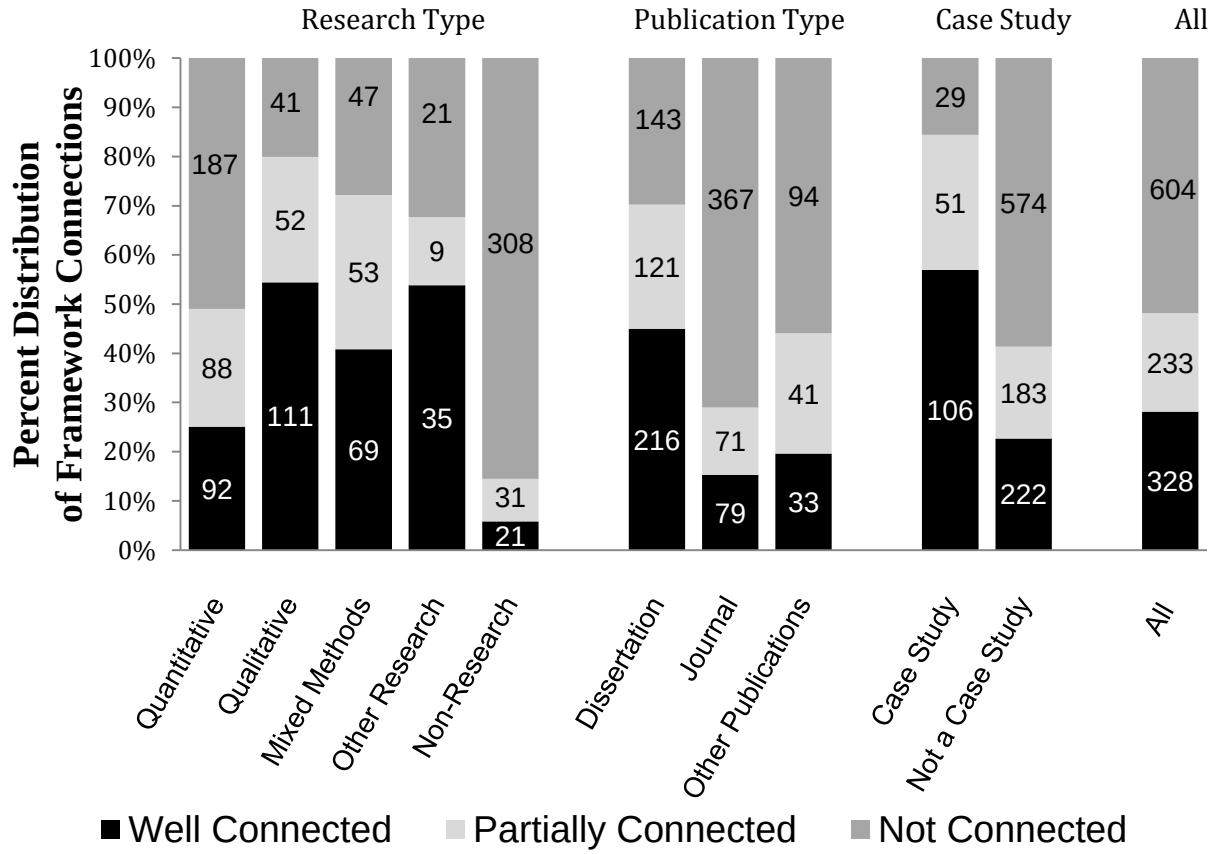

Figure 7. Percentage distribution of papers providing each degree of conceptual framework connections by case study, research and publication type. Numbers within the stacked bars are the number of papers in each subset. *N* = 1,165 papers. Quantitative includes meta-analyses and single-subject research. Qualitative includes action research and design experiments. Other Research includes theory development, literature reviews, and descriptions of technology development for mathematics education. Other Publications includes book chapters, full books, reports, and master's theses.

Limitations

The number of quality indicators for different research types varied, making comparisons difficult. For example, 17 points were possible for mixed methods studies but only five points for non-research papers. This difference in range is problematic, but unavoidable due to the variations in information needed for various types of research. By limiting the measures to the presence or absence of characteristics, we attempted to maximize the balance of weights given to specific characteristics. This choice was also made to maximize our inter-rater reliability and to minimize subjectivity. Limiting to the presence or absence of a characteristic, however, prevented us from measuring how well each characteristic was addressed.. Finally, we recognize that the components within the QF do not necessarily have an equally strong impact on the overall quality of the paper. Without a reliable, valid way to differentiate the effect of each category, however, we chose to weight them as close to equally as possible to avoid inserting personal biases into the metric.

Additionally, specific research types included more components within QF, yielding different numbers of points possible. We therefore interpreted our results with caution and limited our discussions of overall quality to the percentage of quality points earned to avoid bias toward specific research types.

Discussion

The results provide an introductory broad perspective intended to offer an objective view of the quality of the mathematics education technology literature and to initiate a conversation about research quality across the mathematics education community. We envision that further refinement of this process will include some measure of the degree of the appropriateness and/or robustness of each indicator, resulting in stronger measures of quality based on the QF categories. Going beyond the presence or absence of a characteristic would be one way to produce a stronger quality measure. For example, developing more in-depth criteria that capture how conceptual frameworks should be applied throughout the components of a study may provide better guidance to future researchers.

While non-research publications can be quite valuable for researchers and practitioners, these papers contribute a large portion of the mathematics education technology literature base (30.4%) with a relatively low quality (earning an average of only 37% of the possible quality points). We believe that such a proportion of non-research in the literature is inconsistent with the NRC principles of research (Shavelson & Towne, 2002), specifically the third and fourth criteria, using methods that permit direct investigation and providing a coherent and explicit chain of reasoning. In this sample of papers, non-research papers presented strategies for integrating technology into the teaching and learning of mathematics based largely on author experiences and anecdotal evidence. Current literature supplies an abundance of research to support a wide array of instructional strategies. If the field is to move beyond anecdote and opinion, then non-research papers must begin to make connections to the relevant research, particularly when they recommend uses of technology. We do not suggest that non-research papers should be turned into research papers, but that non-research papers should be held to some standards for explicit connections to literature and theory.

We formed two major conclusions from these analyses. First, we found that dissertations accounted for a surprisingly high portion of the literature and research: 39.7% of the available literature and 57.0% of the research studies. As such, nearly half of the available mathematics education technology literature has been produced by the least experienced researchers and contributed the highest average quality. The high quality of dissertations may be due in part to the mentoring and oversight of more experienced researchers, but if so, the reason other publications do not mirror or go beyond the quality of dissertations is unclear. Although space constraints may explain some of the lack of important details in journals, the potential causes for such a divergence are unclear from the available data. From our own experiences developing the present study, we recognize that important details are easy to leave out of a report even though they were attended to. Nevertheless, the high quality of dissertations indicated that emerging researchers are, in fact, equipped with the necessary tools to identify and report key information, which we found to be an encouraging trend.

Second, the overall quality of the mathematics education technology literature is lower than we expected, averaging only 48.9% of the QF points possible. We noted that the quality of research papers, with respect to possible point values (Figure 1) averaged 54.6% over four decades. For mathematics education technology researchers, manuscript reviewers, and editors, these results suggest that more attention is needed on the information being included and excluded from scholarly papers, especially with regard to connections to theoretical frameworks and research designs.

The present study provides a foundation for expanding the discussion of the quality of mathematics education technology literature that will be used to support future research and classroom practice. Interest in technology for mathematics education has grown exponentially over time and will likely continue as technology evolves. Reporting key information in both non-research and research papers is therefore critical for advancing the field. The Quality Framework may be applied beyond the mathematics education technology field. Although its categories were compiled for the present study to specifically examine the quality of mathematics education technology literature, the sources from which the categories were compiled were not specific to the field of mathematics education technology, nor were the criteria within each category. Concerns about the quality of all education research abound (Shavelson & Towne, 2002; Tobin, 2007; Towne, Wise, & Winters, 2005), but explicitly attending to and reporting on the rigorous methods applied to a study may begin to improve its credibility. The application of frameworks such as the QF may help education researchers initiate conversations about quality or provide guidance to authors preparing to present research results or methods for applying research. As education research

becomes more defensible and links between research and practice become more pronounced (i.e., improved reporting of quality indicators), literature resources will become a stronger conduit for helping teachers incorporate the use of strategies and tools, including technology, to improve student outcomes.

References

Abramovich, S., & Ehrlich, A. (2007). Computer as a medium for overcoming misconceptions in solving inequalities. *Journal Of Computers In Mathematics And Science Teaching*, *26*, 181-196.

Burrill, G., Allison, J., Breaux, G., Kastberg, S., Leatham, K., & Sanchez, W. (2002). *Handheld graphing technology in secondary mathematics: Research findings and implications for classroom practice.* Lansing, MI: Michigan State University, Texas Instruments. Retrieved from http://education.ti.com/sites/UK/downloads/pdf/References/Done/Burrill,G.%2520(2002).pdf

Congdon, J. D., & Dunham, A. E. (1999). Defining the beginning: The importance of research design. *IUCN/SSC Marine Turtle Specialist Group*, 4, 1-5. Retrieved from http://mtsg.files.wordpress.com/2010/07/14-defining-the-beginning.pdf.

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey Literature in Meta-Analyses. *Nursing Research*, *52*(4), 256-261. DOI:10.1097/00006199-200307000-00008

Cooper, H. (1998). *Synthesizing research*. Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Creswell, J. W. (2007). *Qualitative inquiry and research design* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2009). *Research design*: *Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.

de Villiers, M. (2004). Using dynamic geometry to expand mathematics teachers' understanding of proof. *International Journal of Mathematical Education in Science and Technology*, *35*(5), 703-724.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L, … Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort* [NCEE 2007-4005]. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20074005/

Easton, J. Q. (2010, May). *Out of the tower, into the schools: How new IES goals will reshape researcher roles*. Presidential session presented at the annual meeting of the American Educational Researcher Association, Denver, CO. Retrieved from http://ies.ed.gov/director/pdf/easton050210.pdf

Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education, 34,* 433-463.

Ellington, A. J. (2006). The effects of non-CAS graphing calculators on student achievement and attitude levels in mathematics: A meta-analysis. *International Journal of Instructional Media, 106*, 16-26.

Fitzer, K. M., Freidhoff, J. R., Fritzen, A., Heintz, A., Koehler, J., Mishra, P., et al. (2007). Guest editorial: More questions than answers: Responding to the reading and mathematics software effectiveness study. *Contemporary Issues in Technology and Teacher Education*, *7*(2). Retrieved from http://www.citejournalorg/vol7/iss2/editorial/article1.cfm

Johnson, S. D., & Daugherty, J. (2008). Quality and characteristics of recent research in technology education. *Journal of Technology Education*, *20*, 16-31.

Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher, 26*, 4-12. DOI: 10.3102/0013189X026007004

Koehler, M. J., Shin, T. S., & Mishra, P. (2011). How do we measure TPACK? Let me count the ways. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 16-31). Hershey, PA: IGI Global. DOI: 10.4018/978-1-60960-750-0.ch002

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lyublinskaya, I., & Tournaki, N. (2011). The effects of teacher content authoring on TPACK and on

student achievement in algebra: Research on instruction with the TI-Nspire™ handheld. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 295-322). Hershey, PA: IGI Global. DOI: 10.4018/978-1-60960-750-0.ch013

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108,* 1017-1054.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2014). *About Mathematics Teacher* [Website]. Reston, VA: Author. http://www.nctm.org/publications/content.aspx?id=9414

Niess, M. L. (2005). Preparing teachers to teach science and mathematics with technology: Developing a technology pedagogical content knowledge. *Teaching and Teacher Education, 21,* 509-523.

Oates, G. (2004). Measuring the degree of technology use in tertiary mathematics courses. In W.C. Yang, S.C. Chu, T. de Alwis, & K.C. Ang (Eds.), *Proceedings of the 9th Asian Technology in Mathematics (ATCM)* (pp. 282-291). Blacksburg, VA: Asian Technology Conference in Mathematics. Retrieved from http://www.any2any.org/EP/2004/2004C178/fullpaper.pdf

Oates, G. (2009). Relative values of curriculum topics in undergraduate mathematics in an integrated technology environment. In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2 pp. 419-426). Palmerston North, NZ: MERGA. Retrieved from http://www.merga.net.au/documents/Oates_RP09.pdf

Özgün-Koca, S. A., Meagher, M., & Edwards, M. T. (2011). A teacher's journey with a new generation handheld: Decisions, struggles, and accomplishments. *School Science and Mathematics, 111*, 209-224. DOI: 10.1111/j.1949-8594.2011.00080.x

Pape, S. J., Irving, K. E., Bell, C. V., Shirley, M. L., Owens, D. T., Owens, S. K., Bostic, J. D., & Lee, S. C. (2011). Principles of effective pedagogy within the context of connected classroom technology: Implications for teacher knowledge. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 176–199). Hershey, PA: IGI Global. DOI: 10.4018/978-1-60960-750-0.ch008

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Peck, C., Cuban, L., & Kirkpatrick, H. (2002). Techno-promoter dreams, student realities. *Phi Delta Kappan, 83*, 472-480.

Rakes, C. R. (2012, February). Research in mathematics educational technology: Study overview. In C. R. Rakes (Chair), *A structured inquiry of research in mathematics educational technology: Findings and implications*. Symposium presented at the meeting of the Association of Mathematics Teacher Educators, Fort Worth, TX.

Ronau, R. N., Rakes, C. R., Niess, M. L., Wagener, L., Pugalee, D., Browning, C., Driskell, S. O., & Mathews, S. M. (2010). New directions in the research of technology-enhanced education. In J. Yamamoto, C. Penny, J. Leight, & S. Winterton (Eds.), *Technology leadership in teacher education: Integrated solutions and experiences* (pp. 263-297). Hershey, PA: IGI Global. DOI: 10.4018/978-1-61520-899-9.ch015

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.; pp. 103-126). New York, NY: Sage.

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

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, D.C.: National Research Council, National Academy Press. Retrieved from http://www.nap.edu/download.php?record_id=10236

Song, F., Hooper, L., & Loke, Y. K. (2013). Publication bias: what is it? How do we measure it? How do we avoid it?. *Open Access Journal Of Clinical Trials*, *5*71-80. doi:10.2147/OAJCT.S34419

Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Teddlie, C., & Tashakkori, A. (2009). *Foundations*

of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences. Los Angeles, CA: Sage Publications.

Tobin, J. (2007). An Anthropologist's Reflections on Defining Quality in Education Research. *International Journal Of Research & Method In Education*, *30*, 325-338.

Towne, L., Wise, L. L., & Winters, T. M. (Eds.). (2005). *Advancing scientific research in education*. Washington, DC: National Research Council, National Academies Press. Retrieved from http://www.nap.edu/download.php?record_id=11112

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: John Wiley & Sons.