2021

Use of Research Tradition and Design in Program Evaluation: An Explanatory Mixed Methods Study of Practitioners' Methodological Choices

Margaret Schultz Patel

Use of Research Tradition and Design in Program Evaluation:

An Explanatory Mixed Methods Study of Practitioners' Methodological Choices

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Margaret Schultz Patel

November 2021

Advisor: Dr. Duan Zhang

Author: Margaret Schultz Patel
Title: Use of Research Tradition and Design in Program Evaluation: An Explanatory Mixed Methods Study of Practitioners' Methodological Choices
Advisor: Dr. Duan Zhang
Degree Date: November 2021

Abstract

The goal of this explanatory sequential mixed method study was to assess whether there were observable trends, associations, or group differences in evaluation methodology by settings and content area in published evaluations from the past ten years (quantitative), to illuminate how evaluation practitioners selected these methodologies (qualitative), and assess how emergent findings from each phase fit together or helped contextualize each other. In this study, methodology was operationalized as research tradition and method was operationalized as research design. For phase one (quantitative), a systematized ten-year review of five peer-reviewed evaluation journals was conducted and coded by journal, research tradition, research design, first author setting, evaluation content area, and publication year. These results were first reported descriptively and then considered for inferential modeling. For phase two (qualitative), interviews, which were informed by the findings that emerged in the quantitative phase, were conducted with a purposive sample of 15 practitioners to gain insight into how practitioners make methodological choices. In phase three (integration), findings were integrated to contextualize emergent learnings from each phase. Evidence of statistically significant associations between research tradition, design, first author setting, and content area were discovered. There were no statistically significant associations observed between either research tradition and publication year or research design and publication year. There was also evidence that evaluations conducted in the quantitative research tradition, as well as experimental designs, were overrepresented in the evaluation literature within the timeframe being reviewed. Finally, this study’s procedures generated a hypothesized grounded theory of how evaluators select methods that provided explanation for phase one findings; this theory should be tested by future researchers.

Acknowledgements

I would like to thank my dissertation committee, including Dr. Duan Zhang, Dr. Antonio Olmos, and Dr. Nick Cutforth, for the ongoing guidance and support. Dr. Zhang, in particular, was instrumental to shepherding me towards completion. I'd also like to thank Dr. Robyn Thomas Pitts for her guidance early on; her thoughts helped shaped this final product. Finally, my husband, pups, friends, and family have been vital in supporting me this process; I will always be grateful.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Compared to the disciplines of economics or social science, for example, evaluation could be considered a relatively new discipline. Therefore, there has still been a need, as some have pointed out in the past, to increase self-knowledge in the field of program evaluation (e.g., Azzam, 2011). For instance, while debates on the merits of various methodological approaches have been rampant and well-documented (e.g. Mertens & Hesse-Biber, 2013; Sechrest, Babcock, & Smith, 1993; Smith, 1994), how frequently each methodological approach has been used in practice or how those approaches were selected has been less well-documented. Similarly, in a somewhat recent survey of evaluators concerning what research on evaluation (RoE) questions they would most like answered, research on methods was one of the most frequently selected topics (Szanyi, Azzam, & Galen, 2013).

Problem Statement

At the time of this review, there were very few, if any, systematic examinations of methodology and methods use or rationale for this use present in the literature. One could certainly find conjecture in the literature about the way evaluation methods were selected, such as, “evaluators…have their favorite evaluation models and methods, usually those in which they were trained” (House, 1994, p. 241). This hypothesized practice of defaulting to favored models and methods, as suggested by House, was problematic because each tradition and approach has been designed to generate a particular type of evidence and answer a particular type of question. A mismatch between method and question could lead to limited utility, accuracy, and validity of evaluation results. It was hypothesized that if the qualitative phase of this study revealed that practitioners did, in fact, select methods based on their preferences and comfort levels, this study would propel the field of evaluation forward by identifying this weakness and prescribing improvements to practice.

Terms of Reference

A distinction should be made between *methods* and *methodology.* Both terms were investigated throughout the course of this study. There were various characterizations of these terms in the literature. For example, in a discussion of how mixed methods have been defined across theorists, Creswell and Plano Clark distinguished methodology as "the process of research," which they suggested includes underlying philosophy, methods, and interpretation of results, while method has been treated as the distinction of whether number or words will be the focus of data collection (p. 2-3, 2018). In this study, methodology referred to the research tradition (such as qualitative, quantitative, or mixed method), while method referred to the type of data collection and analysis strategies used (as discussed by Gliner & Morgan, 2000). In this text, *practitioner* referred to professional evaluators when they were conducting evaluations rather than conducting research on evaluation, developing evaluation theory, or teaching evaluation.

Rationale for Current Study

The current study was conducted for two purposes: first, to interrogate whether there were observable trends or associations in the use of methodologies and methods (through a systematized review of published evaluations) with first author setting, evaluation content area, and publication year; and second, to explore what went into the selection of these by evaluators (through semi-structured interviews with a protocol informed by learnings and questions that emerge from the quantitative phase). At study onset, findings were expected to contribute to the field of program evaluation practice in several ways, including the description of methods trends, description of these methods trends across first author settings and evaluation content areas, the generation of a theory of practitioner rationale when selecting methods, and to illuminate whether certain methods seemed to be unduly privileged by the field.

To begin with, this study systematically collected evidence of trends in evaluation methodologies and methods over a ten-year period, as well as which were most commonly used; there was extremely limited research on this topic present in the literature at the time of study conception. This was problematic because it suggested that practitioners, theorists, and those conducting RoE were unaware of which methods were truly being used most in practice, particularly outside of their own anecdotal knowledge. It would be difficult for a field to move forward when lacking this type of foundational knowledge.

Next, this study examined whether these trends were stable across first author settings and evaluation content areas, as these variables were thought to serve as predictors of methodology and methods; similarly, there were no known investigations of this nature present in the literature at the time of study conception. This gap was problematic as well, since if there was bias lurking in various settings and content areas, this would need to be named and addressed.

Further, this study took an in-depth approach to exploring practitioner rationales for how and why they selected methods; there were no other known studies of this nature present in the literature at the time of study conception. Documenting both the "what" of which methods were being used in the field as well as the "why" was thought to be a crucial step in improving the field, as it would be difficult to improve the practitioner thought process without understanding the thought process in the first place. This author's theorized mechanism of change was that once this thought process was documented, practitioners would become more aware of their own habits and biases, and then gradually learn to select methods better-suited to the evaluation questions at hand. This research was also expected to be useful for theorists, as knowledge about practitioner decisions was expected to provide fodder for future theory development.

Additionally, this study was expected to illuminate the types of evidence unduly privileged by the field. If certain methodologies and methods appeared to be used more often frequently than others, assuming a broad range of content areas and evaluation questions, this could suggest that practitioners were over-relying on select methodologies and methods. These insights were expected to be useful not only to practitioners, but also evaluation educators, evaluation clients, and professional evaluation associations. Understanding these insights was expected to help these stakeholders in the field of evaluation advocate for more systematic, equitable, and pragmatic selection of methodologies and methods. Awareness of these trends and practitioner rationales could encourage evaluation practitioners and commissioners to select their approaches more systematically and appropriately, given evaluation goals and program realities.

Finally, findings from this research were expected to generate recommendations related to new guidelines for credentials in program evaluation (such as those being developed by the American Evaluation Association), particularly if findings were to suggest that practitioners were cherry-picking preferred methods rather than choosing those best suited to each evaluation.

The research questions that were addressed by this study included:

1. Did practitioner use of evaluation methods and methodologies over the past ten years vary by first author setting, evaluation content area, or publication year? (phase 1)

2. How did practitioners describe the process of how they selected evaluations methods and methodologies? How did they describe the thought process for selecting methods and methodologies in light of practical considerations? What factors did they identify as influencing this process (phase 2)?
3. How did practitioners' explanations for how they select evaluation methods and methodologies thematically relate to observed practitioner use of evaluation methods and methodologies? How did these explanations contextualize observed differences, similarities, or associations? (phase 3)?

The first question, which was quantitative, was selected to address an established gap in the literature: there have been few previous investigations focused on use of evaluation methodology and method (one notable exception was Christie & Nesbitt Fleischer, 2010). The goal for the quantitative phase of the research was to document which methodologies and methods were used, as well as to systematically assess whether there were any trends in methodologies and methods used, particularly using the predictors of first author setting, evaluation content area, and publication year.

The second set of questions, which were qualitative, were selected for two purposes. The first purpose was to collect data to develop a theory on how practitioners selected evaluation methods. The second purpose was to strengthen the understanding of the findings of the previous quantitative research question. The overarching goal for this qualitative phase was to explore and contextualize emergent findings from the quantitative phase; specifically, to explore the factors that impacted practitioners' methodological decision-making process and any observed differences or associations. Findings from the quantitative phase were expected to be enhanced by practitioner perspectives, as the interview protocol included questions meant to probe findings from the first phase. In phase 2, interview participants were asked about quantitative trends that emerged in phase 1.

The final, and mixed method set of questions, were selected to integrate the findings of the quantitative and qualitative phases of this study. The goal for this phase was to contextualize the quantitative and qualitative findings in light of each other and ultimately, generate practice improvement recommendations to the field of evaluation.

Chapter 2: Literature Review

Use of Evaluation Methodology

Most of the literature available at the time of this review on use of methodology and methods in evaluation was non-empirical, consisting of reflections, editorials, and discussions (e.g., Norris, 2005; Smith, 1994; Stufflebeam, 2001). A typical example of this body of work was Norris' discussion of how important methodological choice is; Norris asserted that methodological creativity was superior to prescriptive approaches, which were of limited utility in a context-dependent field such as evaluation (2005). Another illustrative example was Stufflebeam's treatise on evaluation methods used in the 20th Century; in this piece, Stufflebeam categorized which methods he felt were worth holding on to and not (2001). While it may be useful to consider the opinions of often famed evaluators, there was little empirical basis for the assertions made in these types of articles.

There were, however, a few empirical examinations of trends in evaluation methodology evident in the literature (e.g., Christie & Nesbitt Fleischer, 2010; Galport & Galport, 2015). For example, following the scientific-based research movement that seemed to be taking off at the time, Christie & Nesbitt Fleischer (2010) conducted what they referred to as a content analysis of three evaluation-focused journals to determine whether there appeared to be a proliferation of randomized controlled trials. Ultimately, they found that non-experimental designs were used most frequently, followed by qualitative and mixed methods designs. This may have been because practical realities dictated that the average program under evaluation was not ready for experimental study (e.g. due to a lack of outcome evidence, data capacity, or newness). Another empirical study closely related to this proposed study concerns trends in Research on Evaluation (RoE) methods (Galport & Galport, 2015). Using a dataset of research on evaluation (RoE) articles published in the *American Journal of Evaluation* from 1998-2014, the authors categorized "methods-focused articles" to "uncover themes and trends in research on evaluation methodologies and techniques" (p. 17). Most relevantly, they found nine themes related to why various evaluation methods were or should be chosen in these RoE articles, including: "multiple units of analysis, maximizing data quality, determining evaluability, measuring fidelity, clarifying theories of change, an emphasis on low-cost or rapid results, a focus on qualitative or mixed methods approaches, and sampling concerns" (p. 24-25). The current study built upon Galport & Galport's research (2015) by cataloging methodologies used, and why they were used, in actual evaluation practice, rather than in RoE.

Methodology Decisions in Evaluation

Similar to the previous research question, much of the published literature on the question of how methodology decisions were made consisted of reflections from practice and editorials (e.g., Braverman & Arnold, 2008; Chelimsky,1998; Chelimsky, 2007; Greene, Lipsey, Schwandt, Smith & Tharp, 2007; Kallemeyn, 2009; Schwandt, 2014, Smith, 1997, Spence & Lachlan, 2010) or prescriptive charges for how these decisions should be made (e.g., Braverman, 2012; Chelimsky, 2012; Mark, 2018; Maynard, Goldstein, & Nightingale, 2016; Julnes & Rog, 2007; Sechrest, Babcok, Smith, 1993).

An illustrative example of this subset of the literature came from Chelimsky (2007), who stated, "From an evaluator's perspective, an a priori judgement about methods without a serious study of the context and specifics of a question is both unsuitable and imprudent in relation to likely evaluation success" (p. 31). Another line of literature that was available on this topic was theoretical or prescriptive. For example, Kundin (2010) provided a framework for how to study evaluators' decisions made in practice that emphasized considering whether evaluators select methodologies based on evaluation theory or if they use their own "practical knowledge," consisting of assumptions, expertise, values, and judgement" (p. 347). Kundin's suggested framework also included considerations of evaluation context and real-time reflection based on changing environments. Similarly, in a discussion piece published in 1994, Chen predicted that in the future, evaluation decisions would be made based on the specific evaluation question under study rather than a dogmatic attachment to quantitative or qualitative methods (Chen, 1994). Similarly, in the same year, House stated, "Originally only quantitative methods were deemed objective enough to be useful for evaluation, which followed beliefs then current in the social sciences...However, we have entered an ecumenical period in which qualitative techniques are seen as legitimate and mixed designs are recommended" (p. 241).

These opinions and suggested frameworks, usually from venerated evaluation theorists or practitioners, were indicative of the type of literature that existed on this question. While this body of work was a useful starting point in documenting method use and how those choices were made or should be made, a more systematic assessment of method use along with further exploration of how those methods are selected would increase self-knowledge in the field.

Conversely, there were a handful of empirical and/or systematic examinations of how methods were selected by evaluators evident in the literature (Azzam, 2010; Azzam, 2011; Christie, 2003; Tourmen, 2009). For example, Azzam (2011) conducted a study that posed several evaluation questions to responding evaluators and asked them to propose designs. Azzam found that design choices were related to methods preferences and reported degree of focus on utility. Alternatively, there did not appear to be associations between design choice and evaluator gender, education, or level of stakeholder involvement in each evaluation scenario. While this study was an important step in pulling back the curtain on how evaluators selected methods, this line of research could be expanded on by examining evaluators' report of how they actually selected methods in previous experiences, rather than positing hypothetically how they might do so.

Similarly, Christie surveyed practitioners about whether they would use theory to inform method selection and found that only 10 percent reported their practice being informed by theory (2003). In a critique of this same research, Datta (2003) asserted that while what respondents purported to do was interesting, an even more useful task would be to review these respondents' evaluation reports to see what they actually do rather than what they say they do. This line of reasoning provided support for the current study. While not exactly empirical, Datta (2007) attempted a somewhat systematic review of federal agency evaluation practice for the purpose of developing policies on method choice. Findings were based on a review of Governmental Accountability Office (GAO) reports, federal regulations, requests for proposals, grants, and reports, Evaltalk discussions (the American Evaluation Association discussion listserv), and personal experience. Ultimately, Datta found that different agencies tended to be inclined toward certain methods, while others were more versatile. These differences seemed to be due to programs lending themselves more naturally to certain designs, agency preferences for one kind of design over another, evaluator training and experience favoring certain methods, and the "politics of methodology" (44). This study sought to expand upon previous methods research in a more systematic and comprehensive manner.

The first line of research that emerged from this systematic review of the literature consisted of opinion-based essays or anecdotal reflections on method use and decision-making. While this body of work may have been instructive, it had limited generalizability or validity for the field. The second line of research that emerged from this review consisted of two past empirical attempts to systematically assess the use of methods in published evaluations and a few surveys of practitioners about how they would hypothetically select evaluation methods. These studies, while more relevant to this currently proposed study, were limited by the following factors; one of these studies was conducted over a decade ago and could stand to be updated; the other was focused on methods used in research on evaluation (RoE) rather than evaluation per se; and the surveys concerned hypothetical situations rather than actual practice. In summary, while there was some literature on the topic of method use and decision-making in evaluation, there has been very limited research or empirical investigation on the topic. This means that there were very limited data available on this subject. The current study was designed to generate empirical data and insights that would build upon and expand these important foundations.

Chapter 3: Methods

This research was a mixed methods study using an explanatory sequential design (Creswell & Plano Clark, 2018). The first phase, which was quantitative, was used to conduct a systematic review of published evaluations to address an observed gap in the literature, as there was limited past research conducting quantitative analysis of observed trends in evaluation methods and methodology. The second phase, which was qualitative, was informed by the findings of phase one. For example, persistent trends that emerged from the first phase were explored in qualitative interviews with practitioners. The findings from phase two were used to explain how and why practitioners chose various methods and methodologies. Finally, the integration phrase allowed for the researcher to weave practitioner rationales for methodology choices together with the observed quantitative trends in practitioner use (see Figure 1). This form of integration was consistent with past definitions of mixed method integration in the sequential explanatory mixed design. For example, Edmonds and Kennedy (2017) noted that:

> The explanatory-sequential approach is a sequential approach and is used when the researcher is interested in following up the quantitative results with qualitative data. Thus, the qualitative data is used in the subsequent interpretation and clarification of the results from the quantitative data analysis...This two-phase approach is particularly useful for a researcher interested in explaining the findings from the first phase of the study with the qualitative data collected during Phase 2. (p.196-197)

This design and integration process was expected to generate unique insights into and recommendations based on observed trends.

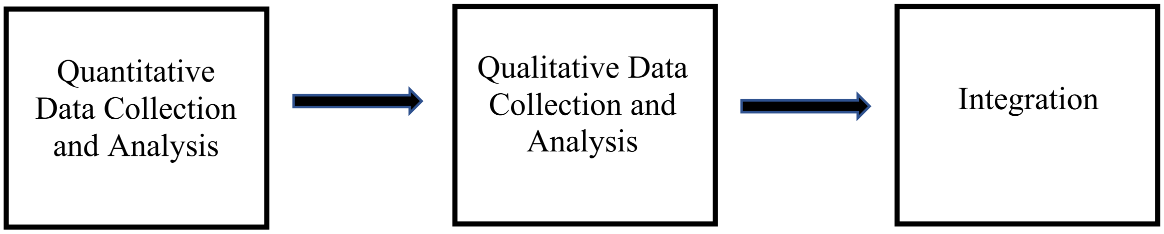
Quantitative Data Collection and Analysis
Qualitative Data Collection and Analysis
Integration

Figure 1: Diagram of Sequential Exploratory Mixed Methods Research

The research questions that were addressed by this study included (also summarized in Table 1):

1. Did practitioner use of evaluation methods and methodologies over the past ten years vary by first author setting, evaluation content area, or publication year? (phase 1)

2. How did practitioners describe the process of how they selected evaluations methods and methodologies? How did they describe the thought process for selecting methods and methodologies in light of practical considerations? What factors did they identify as influencing this process (phase 2)?

3. How did practitioners' explanations for how they select evaluation methods and methodologies thematically relate to observed practitioner use of evaluation methods and

methodologies? How did these explanations contextualize observed differences, similarities, or associations? (phase 3)

Table 1: Research Matrix

Research Question	**Variables**	**Data Sources**	**Data Collection (processes)**	**Data Analysis (products)**
• Did practitioner use of evaluation methods and methodologies over the past ten years vary by first author setting, evaluation content area, or publication year? (phase 1)	Research tradition, research design, year of publication, content area, first author setting	*American Journal of Evaluation, New Directions for Evaluation, Journal of Multidisciplinary Evaluation, Practical Assessment, Research and Evaluation, Evaluation Review*	Systematized review (process)	Database (product)
• How did practitioners describe the process of how they selected evaluations methods and methodologies? How did they describe the thought process for selecting methods and methodologies in light of practical considerations? What factors did they identify as influencing this process (phase 2)?	Practitioner perspective, thought process, and identified contextual factors that contribute to method selection	Purposively sampled practitioners	Semi-structured interview protocol (process)	Interview transcripts (product)
• How did practitioners' explanations for how they select evaluation methods and methodologies thematically relate to observed practitioner use of evaluation methods and methodologies? How did these explanations contextualize observed differences, similarities, or associations? (phase 3)	To be determined (depends on results of first two phases of research)	Data collected in phase one and two of study		

Data Collection

In phase one (quantitative), data were collected through a systematized review of the past ten years of issues of the *American Journal of Evaluation, New Directions for Evaluation, Journal of MultiDisciplinary Evaluation, Practical Assessment, Research and Evaluation, Evaluation Review*. These five journals were selected based on the precedent of past systematic reviews, because they were evaluation-focused, published in English, and had at least ten years of issues available online (e.g., Christie & Nesbitt Fleischer; Coryn, Noakes, Westine, & Schroter, 2007). The number and scope of journals selected was also based on pragmatic considerations of time and resources; ideally, all English language journals from around the world would have been selected, but an approach of that scope was not feasible for this study. While relying on the peer-reviewed literature excluded a substantial portion of evaluation work (such as grey literature or unpublished but utilized work), this pragmatic strategy allowed for a systematic approach. All articles that contained a reference to an evaluation conducted and enough detail to determine the research tradition, at minimum, of the evaluation were included in the study sample. This resulted in 200 articles being selected for coding.

In phase two (qualitative), data were collected through interviews with 15 practitioners using a grounded theory approach. The grounded theory approach was intended to generate a theory of a particular process grounded in the perspective of participants; this mirrored the purpose of this phase, which was to develop a theory of how practitioners select methodology that has naturalistically emerged from participant interviews (Creswell, 2000). The sample size was chosen based on the conventions of a grounded theory approach (i.e., capturing a theory that applies across participants, rather than an in-depth focus on the perspectives or experiences of a few participants). Participants were recruited through the American Evaluation Association (AEA) listserv, professional contacts of the researcher, and snowball sampling from each. To increase the representativeness of the sample, a purposive sample of evaluation practitioners across settings, disciplines, and evaluation content areas were interviewed. To ensure a purposive sample, pre-interview demographic data were collected (See Appendix A). To increase representativeness, interviewees were selected to represent a broad swath of evaluation practitioners in the United States in terms of details related to academic degree/credential, practice setting, years of experience, and field of practice. Semi-structured interview questions were posed in an open-ended fashion to allow for participants to comment without being influenced by the researcher's preconceived thoughts. Interviews included questions related to emergent findings from phase one, as well as questions about the considerations that go into selecting research tradition and design and how these considerations may be influenced by factors such as evaluator training, funders, evaluation purpose, or content area. As indicated by grounded theory, data were collected and coded inductively until a coherent theory started to emerge

(Creswell, 2000). This theory was then tested with participants and settled upon once saturation was reached. The end product was a cohesive theory of how evaluators select evaluation methodologies across settings and content areas.

Phase three (mixed methods) did not require any new data collection.

Data Analysis

In phase one (quantitative), data analysis consisted of both descriptive and inferential analysis. To begin with, the number of articles from each journal that met study criteria and were therefore included in the systematized review were reported in a frequency table. Then, the articles were coded by year, research tradition, research design, first author setting, and content area. These codes were then quantified and operationalized into variable counts (e.g, number of articles using each type of research tradition). To increase the reliability of this coding process, an independent rater was engaged to independently code a sample of articles. Interrater reliability was assessed and found to be sufficient (e.g., complete agreement upon discussion). Then, variable counts were analyzed with descriptive statistics and reported in a frequency table. Finally, a multinomial logistic regression was conducted to explore whether research tradition and design could be predicted by first author setting, evaluation content area, and publication year. This analytic approach was well suited to answering research questions exploring whether there was a predictive relationship between the independent and dependent variables. See Table 2 for an excerpt of the final dataset.

As recommended by previous researchers, the analysis of the data gathered in the systematized review was implemented in a manner intended to maximize trustworthiness, including during the preparation phase, the organization phase, and reporting phase (e.g., Elo, Kaariainen, Kanste, Polkki, Utrainen, & Kyngas, 2014). This included determining the utility of each category contained within each code, considering whether categories were truly distinct, determining the degree of interpretation involved in each categorization, and ensuring that categorizations accurately reflected the information provided by article authors.

Table 2. Quantitative Dataset Excerpt

ID	Journal	Year	Tradition	Design	Content	Setting
1	AJE	2011	Quantitative	Descriptive	Human Services	Higher Ed
2	AJE	2011	Quantitative	Experimental	Education	Higher Ed
3	AJE	2011	Multimethod	None	Other	Higher Ed
4	AJE	2011	Quantitative	Descriptive	Other	Other
5	AJE	2012	Qualitative	Other	Education	Higher Ed

Qualitative data collected in the second phase were inductively coded and analyzed for themes (Creswell, 2012). These data were analyzed using grounded theory procedures and the qualitative software program Atlas.ti. These procedures involved axial coding, revision, deductive coding, and ultimately, theory generation. For the third and final phase, these qualitative data were informed by and combined with previously collected quantitative data.

Researcher Positionality

At the time of this study, this researcher was an evaluator with over ten years of experience as a practitioner. Further, this researcher tended to subscribe to the philosophy that a multi-method or mixed-method approach was generally the most comprehensive and valid. This researcher believed that there was often a mismatch between evaluation questions and methods used, and that evaluators should not shy away from using less familiar methods if they would best serve the evaluation questions under study. This perspective likely influenced the researcher's initial reaction to explanations of method choices, but did not affect final interpretations. Reflexivity journaling was used to minimize this bias.

Ethical Considerations

Given that phase one involved the analysis of secondary data, there were limited ethical concerns for this phase. In phase two, which involved primary data, the rights of research participants were protected through the use of Institutional Research Board (IRB) approval, consent forms, secure data storage, and confidentiality. These processes included COVID-19 protections and protocols.

Chapter 4: Results

This study was comprised of three phases of analysis: quantitative, qualitative, and integration. This chapter presents the results of these three phases in three sections. The first section presents the results of the quantitative analyses. There were two primary purposes of the quantitative analysis in this study; first, to document and describe the methodologies and methods observed in the peer-reviewed evaluation journal literature, second, to assess whether methodologies and methods observed can be predicted by first author setting, evaluation content, and publication year. The second section presents the results of the qualitative analyses. The purpose of the qualitative analysis was to generate a hypothesized theory of how evaluators select methods as well as to contextualize findings from the quantitative analysis. The third section presents the results of the integration of the quantitative and qualitative analyses. The purpose of integration was to integrate the quantitative results with the qualitative results. Ultimately, these analyses were conducted to answer the following research questions:

1. Did practitioner use of evaluation methods and methodologies over the past ten years vary by first author setting, evaluation content area, or publication year? (phase 1)

2. How did practitioners describe the process of how they selected evaluations methods and methodologies? How did they describe the thought process for selecting methods and methodologies in light of practical considerations? What factors did they identify as influencing this process (phase 2)?

3. How did practitioners' explanations for how they select evaluation methods and methodologies thematically relate to observed practitioner use of evaluation methods and methodologies? How did these explanations contextualize observed differences, similarities, or associations? (phase 3)?

Phase 1 Quantitative Descriptive Results

The systematized review of peer-reviewed articles published in five evaluation-focused U.S.-based evaluation journals yielded 200 articles containing explicit mention of an evaluation conducted. Journals were first selected based on inclusion criteria (published in English, North American-based, evaluation-focused) and then searched for relevant articles within those journals. Any articles that mentioned an evaluation conducted and enough methodological detail to at minimum identify research tradition (though ideally, research design as well) were included in the study sample. See Table 3 for a summary of articles included by journal. Articles that met inclusion criteria were mostly well-distributed across the years included in this sample (2011-2020). The number of articles that met inclusion criteria each year generally ranged from 10-25. However, one notable exception was 2020; 42 articles met inclusion criteria in that year. See Table 4 for more details. Articles were coded by research tradition, research design, publication year, first author setting, and content area. Research tradition was operationalized as qualitative, quantitative, or mixed method. Research design was operationalized as type of data collection and analysis strategies used, such as experimental, sequential exploratory, or phenomenology. Both research tradition and research design were coded based primarily on how authors characterized their own evaluation methods; secondarily, if no tradition or design was explicitly identified, the researcher attempted to determine tradition and design based on context clues. A comparison of interrater reliability between the first and second rater initially yielded an agreement of 98.5%. Following discussion, the two raters were able to reach an inter-rater reliability rate of 100%.

By research tradition, the majority of evaluations reviewed were quantitative (52.5%). The next most commonly identified research tradition was mixed or multimethod (34.0%); this proportion of articles represent evaluations conducted in either the mixed method tradition or with both quantitative and qualitative data collected but no discussion of integration or a formal mixed method design ("multiple methods"). Only a small portion (13.5%) of evaluations reviewed represented the qualitative research tradition. See Table 5 for more details.

By research design, the most common research design, across categories, was no design specified (32.0%). This category was open to all three research traditions, but in this sample, only included evaluations conducted in the mixed method and qualitative research traditions. The next most commonly observed designs included experimental designs (22%), quasi-experimental designs (19.5%), and "other" designs (16.0%). "Other" designs were open to all three research traditions, but in this dataset, only included qualitative or mixed methods designs. The least commonly observed research design was descriptive quantitative designs (10.5%). See Table 6 for more details.

By first author setting, evaluation authors tended to be based in traditional higher education institutions (59.5%), followed by research or evaluation-focused firms (21.0%). Firms were defined as groups with more than one full-time staff member. The remaining authors (19.5%) were from various non-research settings such as foundations, government, and independent consultancy. See Table 7 for more details.

Evaluations captured within this sample came primarily from the human service (29.0%) and education (28.0%) fields. Many of the sample studies came from a variety of fields such as criminal justice, international development, and health. Due to small cell sizes, these were collapsed into an "other" category (43.0%). See Table 8 for more details.

Table 3: Evaluations by Journal Article

Journal	**Frequency**	**Percentage of Total**
American Journal of Evaluation	59	29.5%
Evaluation Review	73	36.5%
Journal of Multidisciplinary Evaluation	29	14.5%
New Directions for Evaluation	37	18.5%
Practical Assessment, Research, and Evaluation	2	1.0%
Total	200	100%

Table 4: Evaluations by Year of Publication

Year of Publication	**Frequency**	**Percentage of Total Sample**
2011	17	8.5%
2012	22	11.0%
2013	19	9.5%
2014	16	8.0%
2015	23	11.5%
2016	15	7.5%
2017	16	8.0%
2018	13	6.5%
2019	17	8.5%
2020	42	21.0%
Total	200	100%

Table 5: Evaluations by Research Tradition

Research Tradition	Frequency	Percentage of Total Sample
Quantitative	105	52.5%
Mixed Method/Multimethod	68	34.0%
Qualitative	27	13.5%
Total	200	100%

Table 6: Evaluations by Research Design

Research Design	Frequency	Percentage of Total Sample
Descriptive	21	10.5%
Experimental	44	22.0%
None Identified	64	32.0%
Other	32	16.0%
Quasi - Experimental	39	19.5%
Total	200	100%

Table 7: Evaluations by First Author Setting

First Author Setting	Frequency	Percentage of Total Sample
Traditional Higher Education Institutions	119	59.5%
Research/Evaluation Firm	42	21.0%
Other	39	19.5%
Total	200	100%

Table 8: Evaluations by Content Area

Content Area	Frequency	Percentage of Total Sample
Human Services	58	29.0%
Education	56	28.0%
Other	86	43.0%
Total	200	100%

Quantitative Research Question and Hypothesis

1. Did practitioner use of evaluation methods and methodologies over the past ten years vary by first author setting or content area?

It was hypothesized that there would be associations between each dependent variable (research tradition and research design) and the three independent variables (content area, first author setting, and year of publication). Therefore, to test this hypothesis, chi-square tests of associations were conducted. If associations between research tradition and the three independent variables were statistically significant, it would have been appropriate to further test this hypothesis by conducting a multinomial logistic regression. This regression would be intended to test whether research tradition could be predicted by first author setting, content area, and publication year. Similarly, if associations between research design and the three independent variables were statistically significant, it would be appropriate to further test this hypothesis by conducting a multinomial logistic regression. This regression would test whether research design can be predicted by setting, content area, and publication year.

Chi-Square Tests of Association

Prior to conducting the chi-square analyses, all assumptions of the chi-square test of association were tested and met. Chi-square tests of association were then conducted to determine whether there were any associations between research tradition and content area, author setting, and publication year. These same tests were conducted to determine whether there were any associations between research design and content area, author setting, and publication year. Bonferroni corrections were applied to each test to correct for multiple comparisons. Each chi-square test was conducted with the same sample of 200 articles (n = 200).

Research Tradition and Evaluation Content Area, Author Setting, and Publication Year

Results suggested that there was a moderate, statistically significant association between research tradition and evaluation content area, $X^2(4) = 21.79, p < .001$. Cramer's *V* suggested a moderate effect size (.23). An analysis of standardized residuals suggests that evaluations conducted in the mixed/multiple method research tradition and with a content of other (not human services or education) were a major contributor to the overall chi-square value (standardized residual = 2.5). Conversely, evaluations conducted in the quantitative tradition with a content area outside of human services or education were a very weak contributor to the overall chi-square value (standardized residual = -2.3). There was also a moderate, statistically significant association between research tradition and author setting X^2 (4), = 17.82, $p = .001$. Cramer's *V* suggested a moderate effect size (.21). An analysis of standardized residuals suggests that evaluations conducted in the mixed/multiple method research tradition by practitioners located outside of firms or traditional higher education institutions were a major contributor to the overall chi-square value (standardized residual = 2.4). Conversely, evaluations conducted in the qualitative research tradition by evaluators within firms were a particularly weak contributor to the overall chi-square value (standardized residual = -2.0). Finally, there was not a statistically significant association observed between research tradition and year of publication X^2 (2),= 0.82, *p* =.665. See Table 9 for more details.

Research Design and Evaluation Content Area, Author Setting, and Publication Year

Further, results suggest that there was a moderate, statistically significant association between research design and content area X^2 (8), = 26.35, *p* = .001. Cramer's *V* was moderate, .26. An analysis of standardized residuals suggests that experimental designs conducted within the human services content area were a major contributor to the overall chi-square value (standardized residual = 2.3). Further, evaluations without research designs identified and conducted outside of either the human services or education content area (falling under "other") were another major contributor to the overall chi-square value (standardized residual = 2.2). There was also a weak, statistically significant association observed between research design and author setting X^2, (8), = 21.46, *p* = .006. Cramer's *V* suggested a moderate effect, .23. An analysis of standardized residuals suggests that evaluations without a research design identified conducted outside of a firm or traditional higher education institution ("other"), contributed strongly to the overall chi square (standardized residual = 2.7). Finally, there was no statistically significant association observed between research design and publication year X^2, (4),= 4.60, *p* =.331. See Table 10 for more details.

These results indicated that content and author setting were appropriate predictors for both research tradition and research design in a subsequent logistic regression model; however, publication year would not an appropriate predictor. Therefore, the independent variable of publication year was excluded from the subsequent model. Finally, Bonferroni corrections were applied to correct for multiple comparisons; all statistically significant comparisons remained statistically significant at the .05 level.

Multinomial Logistic Regression

Prior to conducting the multinomial logistic regression, assumptions were tested. Multicollinearity between the independent variables of setting and content was observed. More specifically, the two remaining predictors (first author setting, content area) are statistically significantly associated with each other, with a moderate effect size; this held even with a Bonferroni correction to adjust for multiple comparisons. See Table 11 for more details. While often a composite variable may be developed to address multicollinearity concerns, in this instance, this was deemed inappropriate due to the conceptually distinct nature of the first author setting variable and content area variable. Finally, there was an imbalanced distribution of research tradition and design across first author setting and evaluation content areas among coded articles (see Figures 2-5), which is also problematic for inferential models. Considering the multicollinearity of the independent variables along with this imbalanced distribution, the researcher concluded that this dataset was inappropriate for multinomial logistic regression analyses.

Further, the possibility of nesting between research tradition and research design may have suggested that a multilevel analysis could be more appropriate than logistic regression. However, there were several reasons why this dataset is not appropriate for multilevel analysis. To begin with, there are only two levels, rather than three or more. And, the level of nesting, while theoretically present, is not extensive in this dataset (See Table 12 for more details). Further, simulation research suggests that multilevel analysis with sample sizes less than 50 at level two can lead to biased estimates. (e.g., Maas & Hox, 2005). In this case, research tradition was considered a level two variable; since there were less than 50 observations within the qualitative research tradition, the sample size was insufficient. Therefore, due to lack of levels, the lack of nesting present in the dataset, and the small size of this dataset (N = 200), it was determined that no multilevel analysis would be warranted. Therefore, the quantitative analyses concluded with chi-square tests of association.

Table 9: Significant Associations between Research Tradition and Independent Variables

Comparison	***χ2***	***df***	***N***	***p***	***Cramer's V***
Research Tradition and Content	21.18	4	200	<.001	.230
Research Tradition and First Author Setting	17.82	4	200	.001	.211

Table 10: Significant Associations between Research Design and Independent Variables

Comparison	***χ2***	***df***	***N***	***p***	***Cramer's V***
Research Design and Content	26.35	8	200	.001	.257
Research Design and First Author Setting	21.46	8	200	.006	.232

Table 11: Multicollinearity between Independent Variables

Comparison	***χ2***	***df***	***N***	***p***	***Cramer's V***
Content and First Author Setting	20.49	4	200	<.001	.22

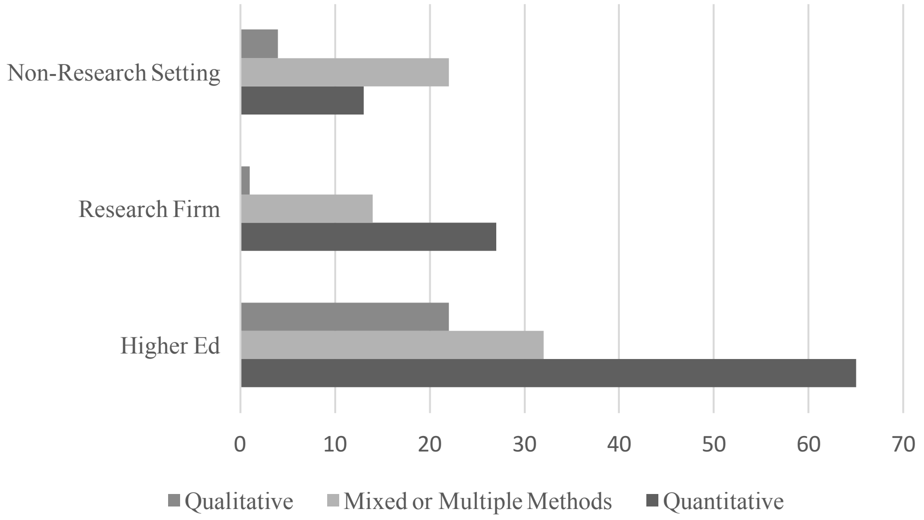

Figure 2: Research Tradition by First Author Setting

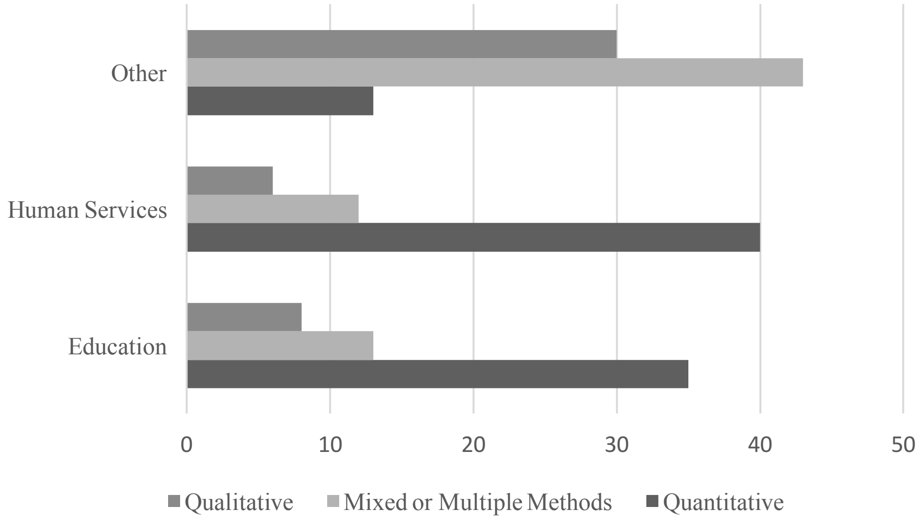

Figure 3: Research Tradition by Evaluation Content Area

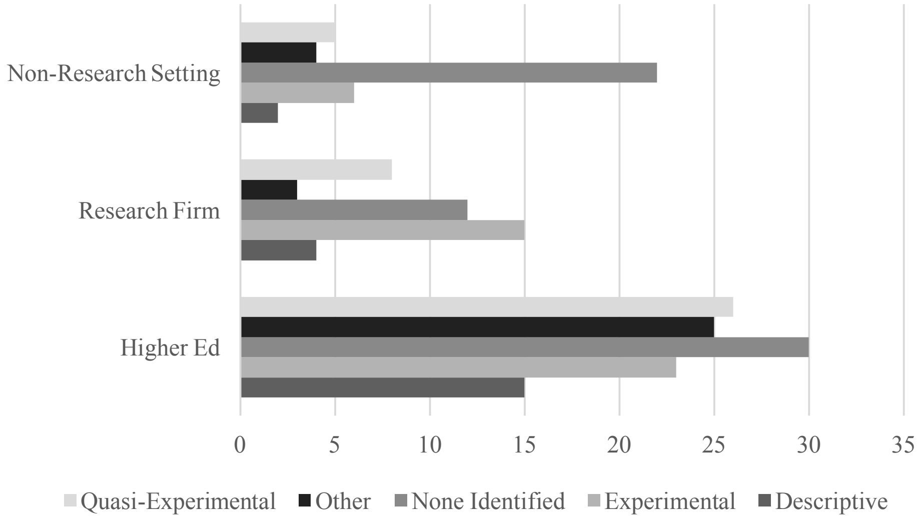

Figure 4: Research Design by First Author Setting

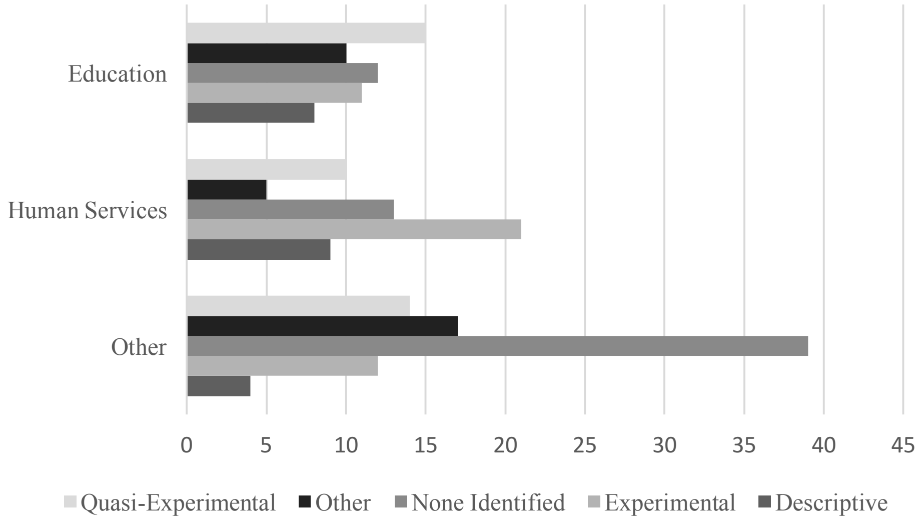

Figure 5: Research Design by Evaluation Content Area

Table 12: Extent of Nesting of Research Design within Research Tradition

Research Design	**MixedorMultiple**	**Qual**	**Quant**
Descriptive	0	0	21
Experimental	0	0	44
None	53	11	0
Other	15	16	1
QED	0	0	39
Total	68	27	105

Quantitative Summary

In summary, articles gathered during the systematic review tended to be quantitative (research tradition) and experimental (research design). The content area of evaluations was quite varied, with the largest proportion of evaluations falling in fields other than human service or education. By first author setting, authors tended to be based in traditional higher education institutions. By year, articles that met inclusion criteria tended to be published in 2020.

There were significant associations between research tradition and author setting, research tradition and content, research design and author setting, and research design and content. There were also significant associations between the independent variables (content and author setting), which implied a level of multicollinearity that was problematic for conducting as a regression. Further, the small sample size and limited nesting observed within this dataset precluded the need for multilevel analyses. Therefore, analyses were concluded with chi-square tests of association.

Phase 2 Qualitative Research Questions

2. How did practitioners describe the process of how they selected evaluations methods and methodologies? How did they describe the thought process for selecting methods and methodologies in light of practical considerations? What factors did they identify as influencing this process (phase 2)?

Phase 2 Qualitative Results

Interviews were conducted with 15 evaluation practitioners who volunteered to participate in interviews. Prior to interviews, participants completed a brief pre-screening survey about the nature of their practice, training, and experience.

Practice

Evaluation practitioners interviewed were based in private research and evaluation firms (33.3%), traditional higher education institutions (26.7%), and non-research settings such as independent evaluators (20.0%), nonprofits/community based service providers, (13.3%), and foundations (6.7%). Further, interviewees reported that they practiced in a diverse range of content areas. See Table 13 for a full listing of responses.

Training

Interviewees mostly held doctorates (40.0%) or were in the process of completing their doctorates (26.7%); a smaller proportion of interviewees held Master's degrees as their highest level of training (33.3%). More specifically, respondents came from a variety of training backgrounds. While about a third obtained an advanced degree in evaluation, (33.3%), the remaining two thirds (66.7%) obtained advanced degrees in fields so varied and disparate, they had to be classified as "other" (example: Economics, Social Work, Literature). Two respondents (13.3%) indicated that they had received post-graduate credentials (in Nonprofit Management and Conflict Resolution; as a Prevention Specialist, respectively).

Experience

Interviewees reported that they had been practicing for an average of 11 years (ranging from 5 to 25 years).

Table 13: Practice Content Responses

Content Area
Behavioral health and prevention
Community educational programming, services provided to investigators.
Diversity in tech
Early childhood education and child welfare organizational health
Economics of Education
Environment and community development
Foundation portfolios
Health
Human services
Human services/education
I do developmental evaluation across topics but most of my work currently is in human services and health
Mental health (specific focus on suicide prevention)
Philanthropy
Poverty, homelessness, workforce development, education
Public Health

Grounded Theory Analysis

Grounded theory analysis procedures as recommended by Cresswell were used (2013). Interviews were conducted until saturation was reached (at 15 interviews). Interviews were first transcribed, and subsequently, open coded. Through the process of open coding, the core phenomenon of how evaluators select methods emerged. Following the identification of this core phenomenon, selective axial coding was conducted to hone in on the nuances of the core phenomenon, as recommended by Cresswell (2013). As recommended by Cresswell, categories such as causal conditions, strategies, intervening factors, and consequences of undertaking strategies were developed through the process of axial coding. Ultimately, this information was used to develop a hypothesized theoretical model of how evaluators select methods (Figure 6). This theoretical model is described in detail below.

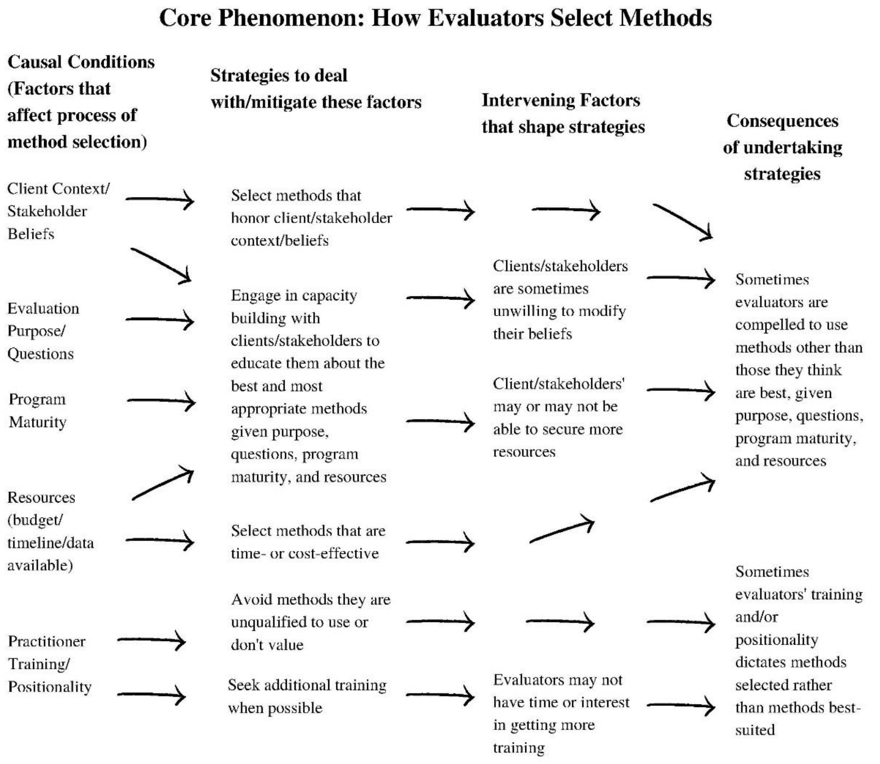
Core Phenomenon: How Evaluators Select Methods
Causal Conditions (Factors that affect process of method selection)
Strategies to deal with/mitigate these factors
Intervening Factors that shape strategies
Consequences of undertaking strategies
Client Context/ Stakeholder Beliefs
Evaluation Purpose/ Questions
Program Maturity
Resources (budget/ timeline/data available)
Practitioner Training/ Positionality
Select methods that honor client/stakeholder context/beliefs
Engage in capacity building with clients/stakeholders to educate them about the best and most appropriate methods given purpose, questions, program maturity, and resources
Select methods that are time- or cost-effective
Avoid methods they are unqualified to use or don't value
Seek additional training when possible
Clients/stakeholders are sometimes unwilling to modify their beliefs
Client/stakeholders' may or may not be able to secure more resources
Evaluators may not have time or interest in getting more training
Sometimes evaluators are compelled to use methods other than those they think are best, given purpose, questions, program maturity, and resources
Sometimes evaluators' training and/or positionality dictates methods selected rather than methods best-suited

Figure 6: Hypothesized Theoretical Model of How Evaluators Select Methods

Causal conditions

Evaluation practitioner interviewees identified several causal conditions, or factors that affect the process of method selection, including client context/stakeholder beliefs, evaluation purpose/evaluation questions, program maturity, resources, and practitioner training/positionality. To begin with, interviewees identified client context as a key factor that affects method selection. Client context relates to the type of organization and what their priorities, beliefs, and values are, particularly related to ways of knowing and historical interface with evaluation methods (e.g, positivist versus constructivist). As one evaluator explained, “A context is really important to me. So... if I'm working in [with indigenous folks] and there's a cultural context to the evaluation, I often want to respect or dig into that a bit...I think cultural context for me is probably the most important thing these days that drives the way I approach things [methods].”

Similarly, practitioners identified stakeholder beliefs, including beliefs, values, and assumptions about different types of evaluation methods, as a key factor that affects how they select methods. For instance, as one interviewee explained, “Perhaps a stakeholder doesn't think that type of information is credible or actionable so, then we have to...kind of do give and take.” As another interviewee mentioned, method selection will depend on “…what the appetite for experimental design, random assignment, etc., is, which, in our work, we found almost no appetite for that.”

Another driving factor evaluators identified as affecting their method selection process was the evaluation purpose and/or evaluation questions. Many practitioners discussed the idea that certain purposes or questions suggest certain methods. An example of this sentiment can be found in these comments by an interviewee: “I definitely like to stretch and play with the different methodologies and that, at the end of the day, any method that my team and I choose for an evaluation is always based on the questions.” Similarly, as another practitioner explained, “the beginning piece that I start with [when considering methods is] understanding what they [the client] really want to know [evaluation purpose].”

Further, several practitioners pointed to program maturity as a causal condition affecting the process of method selection. As one interviewee explained, “It doesn't make sense to be doing a randomized control trial...if you haven't really defined what the ’it’ is that you're implementing.”

Resources, including timeline, budget, and data available, were also identified as a key driver of method selection. Nearly all practitioners interviewed referred to resources as a limiting factor in their ability to select and use ideal evaluation methods. When asked to consider major factors that affect method selection, one practitioner noted, “the timing and budget.” Another participant went into more detail about how time and budget can affect method selection, explaining, “So if I ideally would want to do, like a phenomenological approach…I'd want to do this approach that would make me talk to 12 or 15 people, if you don't have much time or budget, then that's where it's going to get cut down. I'm either going to do a survey or a focus group [instead], or I'm going to [include] fewer people.”

Finally, practitioners frequently acknowledged their own training and/or positionality as having an impact on how they select methods. This training and positionality in turn causes them to select certain methods that they feel comfortable with and trust. For instance, as one practitioner explained the influence of her training, “I was trained in mixed methods at my graduate university, probably with a heavier tilt towards quantitative and more post-positivist thinking." Similarly, another practitioner stated, "I am much more quantitative by skill set...I'd have to go back to school to feel more confident in... choosing [other methods]." Relatedly, positionality, or the practitioner's own values and assumptions certainly seem to play a role in method selection as well. For example, one practitioner explained that, "a mixed methodology is definitely my favorite, my go-to. You can get a whole spectrum of data with that for a wider variety of audiences." Or, as another practitioner stated more simply, "We always use mixed methods." A different interviewee explained this condition in more detail, saying, "I tend to work with groups that have experienced either historical trauma or current trauma because they're exiting prison….and PhotoVoice works really well from a trauma-based perspective with those folks or with indigenous people who have historical trauma…. [this method allows] both storytelling and [is] empowering, and so I like PhotoVoice a lot."

Strategies

Practitioners identified several strategies to either mitigate or work with the causal factors discussed above. The first strategy, perhaps unsurprisingly, in response to client context and stakeholder beliefs about appropriate methodologies, was to select methods that honor client/stakeholder context or beliefs. As one practitioner explained, “usually...in the context in which I'm working,..[my methods] tend to be pretty story-based…[because] it's a storytelling culture.” Similarly, as another practitioner explained, “It kind of goes back to the values of the folks who are in charge of the program, which can help to guide...what methods are appropriate...”

Further, as described above, nearly all interviewees described selecting methods based on evaluation purpose/questions and program maturity. However, as previously discussed, clients/stakeholders often have limiting beliefs or values, which may extend to the types of methods they consider valid or acceptable. In response to this, practitioners frequently described engaging in capacity building to help clients and stakeholders understand why a particular methodology is most appropriate given evaluation purpose, questions, or program maturity. For example, one practitioner described this strategy as, “We take kind of an education approach, when people aren't familiar with [different methodologies]...*”* Or, as another respondent explained, this strategy is about “educating folks [so that they understand why a methodology is more appropriate*.*]” Similarly, a separate practitioner mentioned that, “[certain] types of methodologies usually take some capacity building to [get clients and stakeholders to accept].

Further, in response to resource limitations (including budget, timeline, and/or data available) evaluators described a strategy of selecting methods that are time- or cost-effective. In a response typical of interviewees, one practitioner said simply, “If we have a short turnaround on a due date…[that has an effect on which methods are selected].

Similarly, another practitioner commented, “Funding is a huge consideration in deciding what approach to take...” Providing more detail, another practitioner explained, “I use surveys, because of cost effectiveness. I don't actually like those the best. They're more of a necessity.”

Finally, in response to their own positionality or training, practitioners described avoiding methods they are unqualified to use or don’t value, as well as seeking additional training when possible. As one practitioner explained her positionality, she described sticking to methods she values, saying, “I definitely have a bias, probably, towards qualitative research, and so I probably would pick a qualitative design or a mixed methods [design] with a strong qualitative component, more than people that are more quant-focused.” Similarly, another practitioner revealed a propensity for consistently selecting the same methodology, stating, “We always use mixed methods.” Further, interviewees mentioned defaulting to the methods they have the most experience in, making comments such as, “I have more experience in mixed methods, so I'll admit...that's kind of my preference.” Some practitioners also espoused being willing to seek additional training, making statements such as, “I generally try to not let [my] skill set get in my way...I'm willing to learn new skill sets.”

Note that in this hypothesized substantive grounded theory, multiple causal conditions could influence the strategies selected by practitioners. For example, one practitioner described being open to gaining new skills if an evaluation purpose would be best served by a methodology she was unskilled in, which refers to both the evaluator's training, positionality, and willingness to gain new skills. However, she went on to provide the caveat that if the project timeline was short (e.g, referring to the causal factor of limited resources), she likely would be unable to gain new skills in the allotted time, and would therefore have to default to her original skillset (referring back to training and positionality).

Intervening Factors

Practitioners described several intervening factors that shape the strategies described above. For instance, despite a practitioner's best efforts at education and capacity building, clients/stakeholders are sometimes unwilling to modify their beliefs. In an indicative comment, one interviewee explained that sometimes, despite efforts to education, there is "a funder that says no, you have to go to the certain [methodology]." Or, clients/stakeholders may not be able to secure more sources. An example of this idea is, "I completely [develop an evaluation] scope by virtue of how much money there is to put into [the] evaluation." So rather than identifying the best method based on evaluation purpose or questions, for example, practitioners often have to make pragmatic choices about which methods they can use based on a project budget. Finally, evaluators may not have the time or interest in getting more training needed to pursue unfamiliar methods. In an example of this idea, one practitioner explained this idea, “I have never carried out a full [randomized control trial) by myself and was trained that we don't often do those. So I tend to lean away from true experimental designs.”

Consequences

All of these elements---causal conditions, strategies, intervening factors--come together to yield the following consequences: Sometimes evaluators are compelled to use methods other than those they think are best, given evaluation purpose, questions, program maturity; and sometimes evaluators’ training and/or positionality dictates methods selected rather than evaluation purpose, questions, program maturity.

Qualitative Summary

A grounded theory approach was utilized to illuminate the process of how evaluators select evaluation methods. In interviews, evaluation practitioners described the following causal conditions: client context/stakeholder beliefs, evaluation purpose/questions, program maturity, resources available, and practitioner training/positionality. In order to work around or with these conditions, practitioners described the following strategies: select methods that honor client/stakeholder context or beliefs; engage in capacity building with clients/stakeholders to educate them about the best and most appropriate methods; select methods that are time- or cost-effective; avoid methods they are unqualified to use or don’t value; and seek additional training when possible. Within these strategies, evaluators described several intervening factors that shape these strategies, including: sometimes clients/stakeholders are unwilling to modify their beliefs; clients/stakeholders may not be able to secure more resources; and evaluators may not have time or interest in getting more training. As a whole, these causal conditions, strategies, and intervening factors create the following consequences: sometimes evaluators are compelled to use methods other than those they think are best; and sometimes evaluators' training and/or positionality dictates methods selected rather than the methods best-suited to an evaluation purpose, questions, and program maturity.

Phase 3 Integration Research Questions

3. How did practitioners' explanations for how they select evaluation methods and methodologies thematically relate to observed practitioner use of evaluation methods and methodologies? How did these explanations contextualize observed differences, similarities, or associations (phase 3)?

Phase 3 Integration Results

In this study, integration (phase 3) was achieved primarily by presenting findings from phase 1 (quantitative) to interviewees during phase 2 (qualitative). See Table 14 for a joint display of interview questions that were asked in response to findings from the quantitative phase, as well as the themes of those resultant responses. Further, the grounded theory that emerged from phase 2 may clarify the findings from phase 1. For instance, the overrepresentation of the quantitative research tradition and the experimental research design found in the phase 1 systematic review may be due to the various causal conditions, strategies, intervening factors, and consequence affecting the process of evaluators selecting evaluation methods. For example, perhaps the causal condition of practitioner training or positionality drove practitioners to more frequently select experimental designs. Similarly, the lack of qualitative evaluations observed may also be due to these same causal conditions, strategies, intervening factors, and consequences. In this example, perhaps the causal condition of client/stakeholder beliefs caused practitioners to less frequently select qualitative approaches.

Table 14: Joint Display of Phase 1 and Phase 2 Results

Phase 1 Findings	**Resultant Phase 2 Interview Questions**	**Response Themes**
Evaluations classified as mixed method often did not identify or formal mixed method research design or discuss integration. Very few articles identified under the mixed method tradition contained reference to a formal mixed method research design (such as sequential exploratory).	Are you familiar with the difference between mixed and multi method designs? Are you familiar with formal mixed methods designs?	Many respondents, particularly those trained outside of evaluation-focused programs, were not familiar with the distinction between mixed and multiple methods or formal mixed methods designs. "No, that's not something I've heard." "I think it might be an indicator of that larger problem in research, where mixed methods was considered sort of inappropriate, or illegitimate, by both qual[itative] and quant[itative] researchers."
Very few articles identified under the qualitative research tradition contained reference to a formal qualitative approach (such as phenomenology).	Do you ever use qualitative designs? If so, do you think in terms of a formal qualitative approach?	Some respondents reported taking a mostly quantitative or mixed methods approach. Of those who reported that they do conduct evaluations in the qualitative research tradition at times, most respondents reported that they did not tend to take a formal qualitative approach. "I don't think we honestly are particularly rigorous about selecting a specific qualitative approach, we go pretty high-level. It's...well, we're going to do interviews, we're going to do a focus group."

		“I don't have the luxury of having multiple coders, and having multiple meetings ,and doing the things that are required with [qualitative] research. Sometimes I'm the only coder... and that's how you have to roll. I like to...call it thematic analysis, and leave it at that.” “I kind of think...it has to do with being a nascent field, I think, because...you're [the researcher] getting into formal training and evaluation right, but not very many people do. And certainly not a big proportion of the current evaluation community.”
Articles that met Phase 1 criteria were overwhelmingly in the quantitative research tradition.	In my review, I found that quantitative designs are still the most commonly reported designs in the literature. Why do you think that is?	Respondents had a wide range of opinions on why quantitative designs were most heavily represented in the literature. These opinions ranged from hypotheses that biases in who tends to publish, biases among journal editors, to level of complexity and stakeholder preferences. “So much of evaluation [is] being driven by government and government tends to insist upon quantitative designs.”

It's not surprising in the context of most of the articles being put forward by academics...My guess is that practitioners in the field are probably using a lot more mixed and qualitative methods. But academics are using more quant[itative] methods. And they're the ones getting published.”

“[Quantitative] is easier; it's derivative.”

“I think there’s a deterrent from publishing qualitative [work]. I think qualitative people tend to be more user-... utilization-focused, so they probably care a little less [about publishing].

“I think...evaluation...has followed research, where the gold standard for many is considered a randomized control trial. And it's like well if you can't do that, you could look at some associations and then, if you can't do that...qualitative is...at the bottom.”

Integration of Phase 1 Criteria with Phase 2 Interviewees

The purposive sample in phase two was similar, but not identical, to the sample in phase one. For instance, authors based outside of traditional higher education institutions and research/evaluation firms were more heavily represented in the phase two sample than the phase one sample. Similarly, evaluations focused on the field of human services were more heavily represented in phase two than in phase one; phase one contained more evaluations conducted outside of the field of human services or education. See Table 15 and Table 16 for more details.

A more detailed comparison of data collected in phase one and two demonstrated mostly distinct trends highlighted within each paradigm (quantitative and qualitative). To begin with, data collected in phase one indicated that first authors based in traditional higher education institutions tended to use quantitative methods; those in research or evaluation firms also tended to use quantitative methods; and those based in non-research settings tended to use mixed methods. This contrasted with data collected in phase two, which suggested that those based in traditional higher education institutions tended to use mixed methods; those based in research and evaluation firms tended to use mixed methods; and those based in non-research settings tended to use qualitative methods. See Table 17 and 18 for more details.

There was a mix of agreement and distinctions among data collected in phase one and two related to the content areas of evaluations conducted. Among the phase one sample, first authors based in traditional higher education institutions most frequently conducted evaluations in the field of education, while those in non-research settings and firms most frequently conducted evaluations in the field of human services. These findings were somewhat consistent and somewhat conflicted with data collected in phase two, which revealed that practitioners in traditional higher education institutions tended to conduct evaluations in content areas other than education or human services; those in non-research settings tended to conduct evaluations in human services; and practitioners based in research and evaluation firms tended to conduct evaluations in fields other than education or human services. See Table 19 and 20 for more details.

Table 15: Interviewees by Author Setting

Author Setting	**Frequency in Phase 1 Sample**	**Percentage of Phase 1 Sample**	**Frequency in Phase 2 Sample**	**Percentage of Phase 2 Sample**
University	119	59.5%	4	26.7%
Research/Eval Firm	42	21.0%	5	33.3%
Non-Research	39	19.5%	6	40.0%
Total	200	100.0%	15	100.0%

Table 16: Interviewees by Content Area

Content Area	Frequency in Phase 1 Sample	Percentage of Phase 1 Sample	Frequency in Phase 2 Sample	Percentage of Phase 2 Sample
Human Services	58	28.0%	7*	46.7%
Education	56	29.0%	2*	13.3%
Other	86	43.0%	7	46.7%
Total	200	100.0%	16	106.7%*

*One participant indicated that they focused on the content area of both human services and education.

Table 17: Frequencies of Research Tradition by First Author Setting in Phase One Sample

Research Tradition	First Author Setting			Total
	Firm	Non-Research	Traditional Higher Ed	
Mixed or Multiple	14	22	32	68
Qualitative	1	4	22	27
Quantitative	27	13	65	105
Total	42	39	119	200

Table 18: Frequencies of Research Tradition by Practice Setting in Phase Two Sample

Research Tradition*	**Practice Setting**			**Total**
	Firm	**Non-Research**	**Traditional Higher Ed**	
Mixed or Multiple	3	2	2	7
Qualitative	0	4	1	5
Quantitative	1	0	0	1
Total	4	6	3	13

*There were two missing responses. One respondent reported no preferred or most frequently used research tradition, and one respondent did not answer.

Table 19: Frequencies of First Author Setting by Content Area in Phase One Sample

Content Area	**First Author Setting**			**Total**
	Firm	**Non-Research**	**Traditional Higher Ed**	
Education	7	2	48	57
Human Services	16	12	29	57
Other	18	25	43	86
Total	41	39	120	200

Table 20: Frequencies of Content Area Focus by Practice Setting in Phase Two Sample

Content Area*	**First Author Setting**			**Total**
	Firm	**Non-Research**	**Traditional Higher Ed**	
Education	1	1	1	3
Human Services	2	4	1	7
Other	3	2	2	7
Total	6	7	4	17

*Two respondents indicated that they focused both on education and human services.

Integration Summary

Integration was primarily achieved by incorporating findings from Phase 1 into the interview protocol for Phase 2. While the Phase 2 interview sample was not an exact match for the Phase 1 sample, and there were some conflicting patterns when comparing data collected in each phase, interviewees provided important perspectives on the observed lack of clarity on mixed methods versus multiple methods, mixed methods designs and formal qualitative designs, as well as potential explanations for the overrepresentation of quantitative evaluations in the peer-reviewed literature. Further, the substantive theory of how evaluators select methods that emerged from the grounded theory procedures in Phase 2 provides important context to the quantitative findings in Phase 1; particularly the consequences, which suggest that evaluators' training and positionality sometimes unduly affect the process of method selection, and, due to the firmness of stakeholder beliefs and resource constraints, evaluators are sometimes compelled to use methods other than those they think would be best given evaluation purpose and questions.

Chapter 5: Discussion

This study was conducted to document evaluation methods used (phase one), generate a theory of how practitioners select methods (phase two), and integrate these findings (phase three). Ultimately, these findings were expected to generate important insights and recommendations for the field. Notable findings included associations observed between research tradition, research design, first author setting, and content area; the overrepresentation of quantitative and experimental designs; the lack of clear methods identification, evidence of evaluator training/positionality influencing method selection, and conflicting results between quantitative and qualitative data. The discussion below highlights these insights and makes recommendations to improve the field.

Associations Between Research Tradition, Research Design, Author Setting, and Content Area

The phase one quantitative results provided evidence of associations between research tradition and first author setting, as well as research tradition and evaluation content area. These associations also held true between research design, first author setting, and content area. These associations were intriguing; there were several potential explanations for these findings. One explanation, which was supported by the phase two qualitative results, was that practitioner training and positionality affects the types of evaluations practitioners tended to conduct (both in terms of content area and research tradition). First author setting may also contribute to practitioner positionality (or, positionality could dictate author setting); these factors, in turn, could also influence the types of evaluations practitioners conduct (again, in terms of both content area and research tradition). While many in the field may not find these results or explanations surprising, these findings suggest that evaluators are unduly influenced by their setting, training, positionality, and content area when approaching evaluation methodology, rather than considering evaluation purpose and questions in a balanced manner. This would suggest a need for further training and perhaps, a further setting of standards in the field.

One manner in which this study could inform further training or standard-setting is by disseminating the hypothesized theory of how evaluators select methods developed through the course of this study (developed during phase two). This theory, which illuminated both client and individual barriers to evaluators selecting the most appropriate methods, can be presented as something of a cautionary tale to the unreflective or unexamined evaluator. If evaluators were aware that they are likely to be influenced by their training--which still tends to be within a broader discipline rather than a methods or evaluation program, thereby often assuming the biases of each respective discipline rather than maintaining a neutral stance--and positionality, they may be more cognizant of their own propensities for bias and combat those biases. Similarly, if evaluators were trained on how to mitigate the factors that often come up from the client side (e.g., stakeholder beliefs, resource limitations), they may be better able to lead their clients to the methods best suited for the evaluation questions at hand.

Frequency of Quantitative Research Methods and Experimental Research Designs

Quantitative and experimental design were the most frequently observed research tradition and research design in this study sample. These findings contrast with what Christie and Nesbitt Fleischer reported in their content analysis of three years of evaluation-focused journals, which was that mixed methods were the most frequently observed tradition and non-experimental designs were the most frequently observed designs (2010). There are several possible explanations for these differences. To begin with, since Christie and Fleischer's study sample period was outside of the scope of this study (2004 to 2006 versus 2010 – 2020), this difference could be due to a shift over time. Or, these results could be due to the broader scope of this study. Further, the dominance of experimental designs in the peer-reviewed literature (phase one) seemed to conflict with what practitioners reported they used on a regular basis (phase two). This finding contrasts to what was reported in Christie and Fleisher's content analysis of 3 years of journal publications— non-experimental designs were used most frequently, followed by qualitative and mixed methods designs (2010).

This observed frequency of quantitative research traditions and experimental designs may have been due to a quantitative and/or experimental design bias from journal editors, bias in who publishes in the peer-reviewed literature, or within-practitioner bias, which leads practitioners to be more likely to publish their work when it is quantitative and experimental (as opposed to other tradition designs). This potential disconnectedness between practitioners and the peer-reviewed literature may be creating a false narrative of how evaluation is carried out in practice, and further, may discourage practitioner engagement, particularly among practitioners who do not prioritize the quantitative tradition and experimental designs.

These findings also seem to contradict what Chen predicted (1994). Chen predicted that in the future, evaluation decisions would be made based on the specific evaluation questions under study rather than a dogmatic attachment to quantitative or qualitative methods (Chen, 1994). Similarly, in the same year, House stated, “Originally only quantitative methods were deemed objective enough to be useful for evaluation, which followed beliefs then current in the social sciences...However, we have entered an ecumenical period in which qualitative techniques are seen as legitimate and mixed designs are recommended” (241). Despite these optimistic predictions, this study’s findings suggest that quantitative methods continue to dominate in the peer-reviewed literature, regardless of evaluation question or purpose.

Lack of Clear Methods Identification

It was surprising that a substantial number of articles that met inclusion criteria did not clearly identify or provide enough information for the reader to determine research design, particularly among evaluations conducted in the qualitative or mixed method research tradition. Within the qualitative approach, it is important to know the research design--or approach--that guided the research. Similarly, in mixed methods, it is important to articulate which type of data are being prioritized, or if they truly are being treated equally. It is possible that articles published in evaluation journals, while containing some information about an evaluation that was conducted, are excluding methodological details. This may be because articles published in evaluation journals tend to be less about actual evaluations that were conducted and more about lessons learned from evaluation practice. However, even considering this potential study artifact, it is surprising that article authors are not more explicit with their methodology choices, particularly when considering that the peer-reviewed evaluation literature serves as an opportunity for evaluators to talk to each other (rather, than say, a client who may not understand the technical aspects of methodology). It is the position of this author that the methodological decisions and assumptions that are made by evaluators should be made explicit, so that that reader can make their own judgement about the validity or trustworthiness of the results. For instance, taking a quantitative approach suggests a positivist or post-positivist philosophy, whereas taking a qualitative approach prioritizes a constructivist worldview. These assumptions should be made more explicit to emphasize the ways of knowing that are being prioritized and employed within the field. Not naming these assumptions perpetuates the notion that the quantitative, positivist approach is the only way to conduct research or evaluation. One idea for increasing equity across methods and further professionalize the field is for journals to require a clear identification of both research tradition or design; this would require training for journal editors. Finally, during phase two interviews, several practitioners put forth a theory that since evaluation has its roots as a very applied and pragmatic field, evaluators may think less rigorously about methodologies than those in more academically-minded fields. If so, this also suggests a need for further practitioner training and standards within the field.

Association between Author Setting and Content Area

The associations between the analysis variables are intriguing and warrant further exploration. For instance, the association between author setting and content area is interesting. These results may provide support for the hypothesized theory of how evaluators select methods that emerged from phase 2, particularly the consequence that evaluators' training/positionality often influences them to choose preferred methods rather than the best method to answer a particular evaluation question. To improve as a field, it is the position of this author that high quality evaluators should be basing methodology on evaluation purpose and questions. Using the same tool, no matter what the evaluation question or purpose is, cannot be a best practice. Empowering evaluators to draw from a wide swath of methodological tools when considering method selection may require additional training for many evaluators, particularly those trained within another discipline.

Lack of Consensus on Mixed Methods and Qualitative Research Designs

Based on findings from both phase one and two, there appeared to be a lack of consensus on what mixed methods are, as well as what formal qualitative approaches (or research designs) are. This is problematic, because these inconsistencies across the field render it difficult to make comparisons, build legitimacy, and dialogue across the field. This, again, suggests a need for more training, standards, and perhaps, consensus-building within the field.

Evidence of Evaluator Training/Positionality Influencing Method Selection

This study yielded evidence that as likely expected by many, the process of method selection by evaluators is often highly influenced by evaluator training and positionality. While this is perhaps understandable, it is not ideal for the field. As Chelimsky (2007), states, “From an evaluator’s perspective, an a priori judgement about methods without a serious study of the context and specifics of a question is both unsuitable and imprudent in relation to likely evaluation success" (p. 31). This finding provides additional support for the idea that evaluators need to be aware of this frequent shortcoming so that they can be on guard for it, whether through initial or continuing education.

Conflicting Quantitative and Qualitative Data

Finally, integration conducted during phase three revealed conflicting patterns among the quantitative (phase one) and qualitative data (phase two). While phase one data provided potential evidence of a quantitative bias among practitioners in traditional higher education institutions and firms, that trend was less pronounced in the phase two data. Similarly, while phase one data revealed that first authors based in traditional higher education institutions tended to conduct evaluations in education, phase two data provided evidence that practitioners based in traditional higher education settings tended to conduct evaluations in fields of other than education or human services. There are several potential explanations for these discrepancies. These discrepancies could be due to the vast difference in sample size in each phase ($N = 200$ versus $N = 15$); perhaps more similar trends would emerge if more practitioners were sampled in phase two. Or, it could be that the articles netted through the systematized review of phase one were not representative of each first author's global practice. For instance, perhaps a practitioner published an article containing reference to a quantitative evaluation they conducted, but if asked to identify a preferred method, the practitioner would state a global preference for mixed methods. Finally, these differences may be due to a true difference between the population of practitioners who publish in peer-reviewed journals and the general population of practitioners. Either way, the conflicting findings that emerged from the integration process suggest that there may be some salient distinctions between how evaluation methods are portrayed in the peer-reviewed literature (phase one) and how the average practitioner (phase two) would describe their own process for selecting methods.

Significance of Study

This study made an important contribution the literature because it was the first empirical study on the topic of how frequently research traditions and designs are used in evaluations, as well as the first to generate an empirically-based theory of how evaluators select research traditions and designs. Further, this study was a response to a previous call in the literature to measure methods used in practice rather than asking evaluators to consider hypothetical scenarios (e.g., Datta, 2003). Finally, this study provided extensive evidence of a need for increased training, standards, and formalization of the evaluation field.

Implications

Many of the discussion points presented here suggest a need for increased training, standards, consensus-building, and formalization of the evaluation field. The types of practitioner training that would be beneficial include training on a plurality of methods, formal mixed method designs, the integration aspect of mixed methods, formal qualitative research designs (or approaches), how evaluators’ own biases and positionality may unduly influence their method selection process, and how to mitigate common client-driven factors that lead evaluations away from most appropriate methods. Further, journal editors may also benefit from training or guidance on how their own positionality and bias may be influencing how they select articles to publish, how unduly prioritizing certain traditions and designs may be driving away a large contingent of practitioners, and why it is important to require authors to clearly specify both research tradition and design.

Further, this study’s results suggest that evaluation-focused journals should be more rigorous in their requirements for research design specification in published articles. This should be a basic requirement of formal, academic, peer-reviewed publications. If the reader cannot discern the research design of an evaluation being presented, there are important assumptions and contexts that are not being revealed. Similarly, the lack of common language likely fuels the divide between research traditions. For example, if qualitative evaluators are publishing in the same journal as quantitative evaluators but avoid using the formal identification of their methods, they may never be taken seriously those who skew towards the quantitative side of the spectrum.

Additionally, in order to become well-versed in formal research designs, many evaluators may need additional training. Several practitioners interviewed who identified themselves as specializing in mixed methods or qualitative methods were unfamiliar with any formal research designs within those traditions. While clients may not be interested in these formalities, it is crucial that practitioners understand the implications and underpinnings of their methodological choices. And while many practitioners may understand these intuitively without having the formal language to describe them, the more evaluators opt out of assuming the formality of more traditional disciplines, the more slowly the field will gain broad acceptance.

Further, to address the issue of evaluators becoming unduly influenced by their own (non-evaluation) discipline, training, and positionality, evaluation students should be presented with the grounded theory of how evaluators select methods developed through this study. Perhaps as discipline-neutral evaluation programs (i.e., methods programs) become more prevalent, these programs should be held in higher regard than evaluators who receive a degree in psychology, for example, and then apply their social science skills to evaluation. Practitioners need to be aware that their own training and positionality can unduly influence them. They need to pursue methodological plurality so that they are best equipped to answer a range of questions. They need to be aware of the common factors that often come up when working with clients, and strategies for how to possibly mitigate these factors. Overall, evaluators seeking to improve their practice would be well-advised to engage in reflective practice that interrogates their own assumptions, biases and positionality to ensure they are not leaving out any crucial perspectives or doing their clients a disservice by using their own preferred methods.

Limitations

There were several limitations within this study. To begin with, relying on the peer-reviewed literature to assess evaluation methods being used in practice was an intrinsically flawed approach, since the peer-reviewed literature is likely a very skewed sample of the body of evaluations being conducted. Further, the number of interviews conducted may be considered a limitation; grounded theory procedures generally require a larger sample size. Additionally, there was not an exact match in the sampling frames from phase one to phase two. For example, among evaluations met inclusion criteria, the majority of authors were based in traditional higher education institutions. However, the practitioners who volunteered to be interviewed during phase 2 were mostly based at research and evaluation firms and non-research settings, such as independent evaluators or within nonprofits. These limitations may reduce the external validity of this study. Further, the observational and non-probability-based sampling used in this study reduced the internal validity of this study. Finally, as a mixed method study, it would have been useful to more closely mirror the data collection in each phase so that data collected in each phase could be compared and integrated more consistently. For example, it may have been helpful to be able to compare practitioner training and positionality in each phase. While training and positionality data were collected in phase two, these data were not collected in phase one. Though this type of integration was not the original intent of the study, mirroring data collection processes may have yielded important additional insights related to the research questions of this study.

Directions for Future Research

Future research should seek to build upon this line of research by generating a much larger sample size so that these associations could be explored at a higher, and more predictive, level. This could involve including more journals or reviewing additional years of journals. Further, it would be useful to assess methodologies and methods used outside of the peer-reviewed literature. Similarly, future research should endeavor to gather a broader cross-section of practitioners to test whether the hypothesized grounded theory of how evaluators select methods applies to additional populations of practitioners. Finally, future researchers should consider developing a mixed method study on this topic of method selection that allows for more direct comparison between quantitative and qualitative phases; this line of research could contribute to a more nuanced understanding of how practitioners select methods, as well as to help inform recommendations for how to improve this process.

Conclusion/Summary

This study found that there were associations between research tradition, design, author setting, and content area. There was no evidence that traditions or designs are associated with publication year, which may suggest that methods do not come and go out of fashion at certain time points. This study also found that evaluations conducted in the quantitative research tradition, as well as experimental designs, were overrepresented in the evaluation literature, especially compared to what the sample of practitioners reported they use on a regular basis. Finally, this study generated a hypothesized grounded theory of how evaluators select methods that provided a potential explanation for the phase one findings; this theory should be tested by future researchers. The findings from this study should be utilized to inform evaluator training, standards, and professionalization of the field. Understanding the insights revealed in this study should help stakeholders in the field of evaluation advocate for more systematic, equitable, and pragmatic selection of methods, rather than defaulting to preferred methods. Additionally, this research should be useful for theorists, as knowledge generated in phase two about how practitioners select methods should inform future theory development. Further, awareness of method trends observed and practitioner rationales should encourage evaluation practitioners and commissioners to select their approaches more systematically and appropriately, given evaluation goals and program realities. Finally, findings from this research generated recommendations that could be used for informing the development of credentials in program evaluation (such as those being developed by the American Evaluation Association), since these findings do suggest that sometimes practitioners are cherry-picking preferred methods rather than choosing those best suited to each evaluation.

References

Azzam (2010). Evaluator responsiveness to stakeholders. *American Journal of Evaluation, 31*(1) 45-65.

Azzam, T. (2011). Evaluator characteristics and methodological choice. *American Journal of Evaluation, 32*(3), 376 - 391.

Braverman, M. T., & Arnold, M. E. (2008). An evaluator's balancing act: Making decisions about methodological rigor. In M. T. Braverman, M. Engle, M. E. Arnold, & R. A.

Rennekamp (Eds.), *Program evaluation in a complex organizational system: Lessons from Cooperative Extension. New Directions for Evaluation, 120*, 71–86.

Braverman, M. (2012). Negotiating measurement: methodological and interpersonal considerations in the choice and interpretation of instruments. *American Journal of Evaluation, 34*(1), 99-114.

Cordray, D. (1993). Synthesizing evidence and practice. President's address. *Evaluation Practice, 14(*1), 1992, 1-8.

Chelimsky, E. (1998). The Role of Experience in Formulating Theories of Evaluation Practice. *American Journal of Evaluation, 19*(1), 35–55.

Chelimsky, E. (2007). Factors influencing the choice of methods in federal evaluation practice. *New Directions in Evaluation, 113*, 13-33.

Chelimsky (2012). Balancing theory and practice in the real world. *American Journal of*

Evaluation 34(1) 91-98.

Creswell, J. & Plano Clark, V. (2018). Designing and Conducting Mixed Methods research (3rd edition). Los Angeles: Sage Publications.

Chen H. (1994). Current trends and future directions in program evaluation. *Evaluation Practice, 15*(3), 229 - 238.

Coryn, C., Noakes, L., Westine, C., & Schroter, D. 2011. A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*(2), 199-226.

Christie, C. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation, 97*, 7-35.

Christie, C. & Nesbitt Fleischer, D. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation 31*(3) 326-346.

Cresswell, J. (2000). Qualitative Inquiry and Research Design: Choosing Among the Five Approaches (1st edition). Los Angeles: Sage Publications.

Datta, L. (2003). Important questions, intriguing method, incomplete answers. *New Directions for Evaluation, 97*. 37-46.

Datta, L. (2007). Looking at evidence: what variations in practice might indicate. *New Directions for Evaluation*, *113*, 35-54.

Elo, S., Kaariainen, M., Kanste, O., Polkki, T., Utrainen, K. & Kyngas, H. (2014). Qualitative content analysis: A focus on trustworthiness. *SAGE Open, 4(*1).

Galport, M., & Galport, N. (2015). Methodological trends in research on evaluation. In Paul R. Brandon (Ed.), Research on evaluation. *New Directions for Evaluation, 148,* 17–29.

Gliner, J. A., & Morgan, G. A. (2000). *Research methods in applied settings: An integrated approach to design and analysis.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Greene, J., Lipsey, M., Schwandt,T., Smith, N., & Tharp, R. (2007). Method choice: Five discussant commentaries. *New Directions for Evaluation, 113,* 111-127.

House, E. (1994). The future perfect of evaluation. *Evaluation practice, 15*(3), 239-247.

Julnes, G. & Rog, D. (2007). Current federal policies and controversies over methodology in evaluation*. New Directions for Evaluation, 113*, 1-12.

Kallemeyn, L. (2009). Methodological changes and respecting stakeholder dignity. *American Journal of Evaluation, 30*(4) 575-580.

Edmonds, W.E, & Kennedy, T.D. (2017). *An applied guide to research designs: Quantitative, qualitative, and mixed methods* (2nd ed). Sage Publications.

Kundin, D. (2010). A conceptual framework for how evaluators make everyday practice decisions. *American Journal of Evaluation, 31*(3), 347-362.

Spence, P. & Lachlan, K. (2010). Disasters, crises, and unique populations: suggestions for survey research. *New Direction for Evaluation, 126*, 95-106.

Maas, C. & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology,1*(3), 86–92.

Mark, M. (2018). Strengthening links between evaluation theory and practice, and more: Comments inspired by George Grob's 2017 Eleanor Chelimsky Forum Presentation. *American Journal of Evaluation, 39*(1) 133-139.

Maynard, R., Goldstein, N., & Nightingale, D. S. (2016). Program and policy evaluations in practice: Highlights from the federal perspective. In L. R. Peck (Ed.), Social experiments in practice: The what, why, when, where, and how of experimental design & analysis. *New Directions for Evaluation, 152,* 109–135.

Norris (2005). The politics of evaluation and the methodological imagination. *American Journal of Evaluation, 26*(4), 584-586.

Sechrest, L., Babcock, J., Smith, B. (1993). An invitation to methodological pluralism. *Evaluation Practice, 14(*3), 227-235.

Smith, M.F. (1994). Evaluation: review of the past, preview of the future. *Evaluation practice, 15*(3), 215-227.

Schwandt, T. (2014). On the mutually informing relationship between practice and theory in evaluation. *American Journal of Evaluation, 35*(2), 231-236.

Smith, N. (1997). Functions of the evaluation proposal in preordinate and emergent studies. *Evaluation Practice, 18*, (f), 1997, 17-24.

Stufflebeam, D. (2001). Evaluation Models. *New Directions for Evaluation, 89*, 7-98.

Stufflebeam, D. L. (2016). Factors that influenced my conduct of evaluations and evaluation training programs. In D. D. Williams (Ed.), Seven North American evaluation pioneers. *New Directions for Evaluation, 150*, 41–49.

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evaluation, 30*(1), 7-30.

Appendix A: Pre-Interview Screening Survey

1. How many years have you been practicing as an evaluator?

2. What is the primary setting in which you practice evaluation?

a. University/Traditional Higher Education Institution

b. Private Research/Evaluation Firm

c. Nonprofit/ Community-Based Service Provider

d. Government

e. Other (Please fill in:______________)

3. What is the primary topic you evaluate (e.g., human services, public health, education, economics, etc.)?

4. What state are you practicing in?

5. What is your highest academic agree?

6. What subject is your degree in ?

7. Were there any field practice requirements in this degree program? If so, please describe:

8. Have you earned any post graduate credentials? If so, please describe:

Appendix B: Phase Two Semi-Structured Interview Questions

Original Questions

1. How do you usually go about selecting evaluation methods? For the purposes of this study, methods refer to research tradition (quantitative, qualitative, mixed method) and research design (ex: experimental, grounded theory, explanatory).

2. What factors influence your decision? Do those factors seem to have more or less of an influence on you at certain times?

3. Do you have a few favorite methods? Do you tend to use those more often than other methods?

4. What is the ideal situation for selecting methods?

5. How often do you publish evaluation results in peer-reviewed journals?

Additional Questions (based on initial findings)

6. Are you familiar with the difference between mixed and multi method designs?

7. Are you familiar with formal mixed methods designs?

8. Do you ever use qualitative designs? If so, do you think in terms of a formal qualitative approach?

9. In my review, I found that quantitative designs are still the most commonly reported designs in the literature. Why do you think that is?

10. Preliminary results suggest that research tradition is associated with author setting and content area. Do you find this surprising? Why do you think this is?

11. Are there any other thoughts you have about this topic?