INFORMING EVALUATION PRACTICE THROUGH RESEARCH ON EVALUATION

by

Zhi Li

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Research, Measurement, and Evaluation

Charlotte

2024

Approved by:

_____
Dr. Carl Westine

_____
Dr. Chuang Wang

_____
Dr. Xiaoxia Newton

_____
Dr. Kelly Anderson

ABSTRACT

ZHI LI. Informing Evaluation Practice through Research on Evaluation
(Under the direction of DR. CARL WESTINE)

This dissertation advances research on evaluation (RoE) through a trio of studies focusing on the role of context and the innovative use of Linguistic Inquiry and Word Count (LIWC) software in formative evaluation in a qualitative research project. The first article extends Coryn et al. (2017) and elucidates how various contextual dimensions—evaluator, stakeholder, organizational/program, and historical/political—impact the quality and outcomes of evaluations. It underscores the intricate relationship between these dimensions and the evaluation process. The second study leverages LIWC to scrutinize the potential interviewer effects on data collection quality, particularly focusing on the authenticity and emotional tone of interview responses. It intriguingly finds that the demographic alignment between interviewer and interviewee does not significantly alter these LIWC summary variables, challenging assumptions about demographic influences on data quality. The third article expands the application of LIWC to identify linguistic patterns that signal the richness of data, aiming to refine data collection methodologies. This article advances formative evaluation techniques by demonstrating how LIWC can uncover nuanced linguistic indicators of data quality. Collectively, this dissertation highlights the critical role of contextual understanding in RoE and establishes LIWC as a formidable tool for improving the ethical and effective evaluation of qualitative research. The dissertation advocates for a nuanced, context-aware, and technologically informed approach to evaluation that promises to elevate the standards and efficacy of formative evaluation of qualitative research.

ACKNOWLEDGEMENTS

I am incredibly grateful to Dr. Carl Westine, my academic advisor and dissertation chair. His unwavering support, enlightening guidance, and keen insights inspired me to dive deeper and broader into my research, transforming my dissertation into something far beyond my initial vision. Without his guidance, I simply couldn't have reached the finish line.

A heartfelt thank you goes to my dissertation committee—Drs. Chuang Wang, Xiaoxia Newton, and Kelly Anderson. Their discerning feedback and precious advice were pivotal in shaping my work. Their patience and generosity in sharing their knowledge have been invaluable to me.

I also owe a massive thank you to the ERME program faculty—Drs. Sandra Dika, Claudia Flowers, Stella Kim, Rich Lambert, and Jae Hoon Lim. They've been not just teachers but true pillars of support, guiding me from a tentative beginner to a confident researcher. To my friends in the ERME program—Kristin, Qiao, Ting, Tong, Tuba, and Yi—your friendship, support, and shared wisdom have largely improved my learning and growth.

Most importantly, I want to give my biggest thanks to my family. My parents, Hua Li and Jing Liu, have given me nothing but unconditional love and support. And to my cat, Pipi, thank you for being my constant companion. They have been my anchor, giving me the strength to push through and complete my dissertation.

DEDICATION

This dissertation is dedicated to my parents, Hua Li and Jing Liu, whose unwavering support and endless encouragement have been my guiding lights.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RoE	Research on Evaluation
LIWC	Linguistic Inquiry and Word Count
CIPP	Context, Input, Process, and Product
NSF	National Science Foundation
EFA	Exploratory Factor Analysis
PAF	Principal Axis Factoring
PCA	Principal Component Analysis
AI	Artificial Intelligence

INTRODUCTION

Numerous evaluation scholars and theorists have noted the importance of investigating evaluation theories, methods, and practices for its possible benefits. This process, named as "research on evaluation" (RoE), was defined as "any purposeful, systematic, empirical inquiry intended to test existing knowledge, contribute to existing knowledge, or generate new knowledge related to some aspect of evaluation processes or products, or evaluation theories, methods, or practices" (p.161) by Coryn et al. (2016). This dissertation aims to contribute to this literature through a set of empirical RoE studies. According to Mark (2008), RoE is the conduit for helping evaluators select and defend evaluation practice decisions from a multitude of options. In this vein, the intention is to improve evaluators' decision-making concerning various aspects of evaluation practice. Thus, this dissertation first takes stock of what has been learned from the body of recently published RoE by synthesizing the findings of these RoE studies and empirically examining the techniques evaluators use for assessing evaluation context. Previous efforts, notably by Coryn, et al. (2017), fall short in providing this level of detail. Next, it turns the attention to the critical need for evaluators to ground their decision-making using evidentiary practices (Coryn et al., 2016; 2017). To further inform prescriptive practice, it examines the viability of using an automatic text analysis tool, Linguistic Inquiry and Word Count (LIWC), to help monitor and formatively evaluate qualitative data collection processes. Present measures were used to test for interviewer effects and also explore how evaluators can use the tool for factor analysis to improve their practice. Collectively, this dissertation initiates a data analytics-based strategy to assist evaluators in their examination of the evaluation context and conducting formative evaluation.

Starting from the 1970s, studies on evaluation began appearing in published literature (e.g. Bernstein & Freeman, 1975; Patton et al., 1977; Weiss, 1977). These studies primarily focused on RoE utilization (e.g. Cousins & Leithwood, 1986; Leviton & Hughes, 1981), leading to the development of influential theories and refined approaches to prescribe and define practices. After a while, the momentum surrounding RoE waned, and contributions to the RoE library seem to have declined until the beginning of the 21st century (e.g. Alkin, 2003; Henry & Mark, 2003). However, since Christie's (2003) call for more RoE, it has regained scholars' and theorists' attention (Alkin, 2003; Coryn & Westine, 2015; Hansen et al., 2013; Mark, 2008; Vo, 2013), and the number of RoE studies has steadily grown over time (Coryn et al., 2017; Johnson et al., 2009; Vallin et al., 2015; Webb et al., 2017). However, although interest in RoE is growing, the evaluation field still lags well behind other social sciences in the pursuit of developing an evidence base to inform practice.

New and better RoE studies bring many possible benefits to evaluation practice by providing prescription and guidance for practitioners (Mark, 2008; Smith, 1993). In recent RoE studies, some issues have been addressed more frequently with this practical intention, such as evaluation utilization, ethics, and the development of frameworks (Milzow et al., 2019; O'Connor, 2023; Szanyi, 2013). These studies highlight the challenges of navigating evaluation use, which echo previous scholarly attention to the matter (Cousins et al., 2004). Evaluation use requires the careful involvement of stakeholders as both the evaluation competency and decision-making of all parties affect the overall evaluation use (Cousins & Leithwood, 1986; Johnson et al., 2009). Yet, evaluators also face many ethical challenges due to stakeholder involvement (Gedutis et al., 2022; Morris & Cohn,1993). For example, evaluation findings may conflict with stakeholders' expectations, resulting in attempts to have evaluators modify their findings, and

potential negative outcomes could impact the role of stakeholders and their own willingness to be supportive of the evaluation effort. Cultural competency is also an important factor that intersects with both evaluation use and ethics. Chouinard and Cousins (2009) recommend that evaluators create measures tailored to specific cultures, enhance evolving cultural concepts, and concentrate on the relationships between evaluators and stakeholders. In turn, better and closer relationships with the stakeholders and the community will lead to the encouragement of participation in the evaluation, which in turn leads to better evaluation use.

RoE contributes to our understanding of the fundamental issues that exist in evaluation (Mark, 2008). It helps evaluators link theory to practice, assists evaluators in understanding various evaluation contexts, and guides evaluators to work in real-world situations (Szanyi et al., 2013). RoE studies that focus on evaluation activities help evaluators develop insight into the ethical challenges they may face while working in real settings. From there, it is important for the field to plan and test new strategies for how to deal with these kinds of issues.

This dissertation is structured as a three-article dissertation of RoE and aims to bridge the theory-practice gap further. The first paper provides necessary insight into the process of understanding evaluation context by extending Coryn et al. (2017) to synthesize findings of RoE context. Additionally, papers two and three explore the use of data analytics to help evaluators judge project-based qualitative data collection efforts. The second paper attempts to legitimize the use of the LIWC software to efficiently monitor and formatively evaluate interview data collection processes of a qualitative research project using the software's built-in summary variables. The third paper utilizes the research team's own perspectives on the value of data richness with respect to the interview data and examines its factor structure using the full scope of LIWC variables. By using data analytic tools, this dissertation shows how evaluators can

systematically and diagnostically use LIWC to empirically assess the internal values of project teams and formatively evaluate project team data collection practices.

In summary, findings from the set of three RoE studies help evaluators assess the evaluation context and provide strategies to evaluate qualitative data collection efforts ethically and efficiently. Each of the three studies is sequentially described in greater detail below.

Overview of Study: The Three Articles

The First Article

The first article serves as a quasi-needs assessment derived from the recent RoE literature. Its purpose is to understand what recent empirical RoE context reveals about the influence of context in evaluation studies. The study builds on Coryn et al. (2017) to show what has been studied related to the RoE context and what has been learned from it.

Using a qualitative deductive coding method, 58 research articles identified by Coryn et al. (2017) and 14 peer-reviewed scholarly RoE context articles published from 2015 to 2019 were coded into NVivo 12 and analyzed. Guided by the Vo (2013) framework, relevant article paragraphs were coded for each context dimension: evaluator, stakeholder, organizational, program, historical, and political. Coded segments were then coded for descriptors, which were analyzed to identify the relationship between each descriptor and the influence on evaluation practice. This study highlights the significance of understanding the evaluator, stakeholder, organizational/program characteristics, and historical/political contexts within evaluation research. It emphasizes the necessity of a nuanced approach to evaluation, informed by these four dimensions. Aligning with previous research, our findings confirm that context crucially influences evaluation processes and outcomes. The study advocates for flexible, inclusive, and systematic evaluation methods tailored to the specific context of each project. For practitioners,

it suggests the importance of thoroughly considering each project's unique context to shape evaluation efforts effectively.

The Second Article

The second article aims to explore the use of LIWC to assist evaluators in monitoring the data collection process and identifying high-quality data when evaluating a research project. Specifically, we utilized LIWC-22 to test interviewer effects, which can be important criteria in an evaluation with a focus on data collection and tracking of the research process. This study draws upon the example of an evaluation of a qualitative research project involving many successive interviews by a small team of interviewers. Guided by Stufflebeam's Context, Input, Process, and Product (CIPP) evaluation model (Stufflebeam, 1971), this study systematically examines various components of the data collection process with the goal of assisting the research team in improving their efforts and adding credibility to their research findings.

In this study, we explore whether LIWC can effectively assist evaluators in monitoring the data collection process and identifying instances where data collection quality is perceived to be inconsistent. We use LIWC-22 to test interviewer effects in the two LIWC summary variables, authenticity and emotional tone, which are important criteria in an evaluation that focuses on data collection and tracking the research process.

In this study, authenticity and emotional tone scores were analyzed as dependent variables, influenced by several independent variables: the type of question posed by interviewers, the race/gender of the interviewers, and the degree of demographic congruence (such as race and gender) between interviewer and interviewee. Through a two-way ANOVA, significant findings emerged: the nature of the question significantly affected both authenticity and emotional tone. Furthermore, a notable interaction was found for authenticity between the

alignment of race between interviewer and interviewee and the type of question asked. Emotional tone was significantly influenced by the interaction between the interviewer's race/gender and the type of question. However, no significant main effect was observed for the alignment of interviewer and interviewee demographics on either authenticity or emotional tone. These results underscore the effectiveness of using LIWC-22 to assess and inform the data collection stage in qualitative research projects.

The Third Article

The third article extends the second article and aims to achieve two objectives. The first one is to understand better the construct of the human judgments of the qualitative interview data. The second one is to demonstrate the potential of using LIWC to aid evaluators in conducting formative evaluations. Drawing on the same evaluation context as the previous article, we conducted a formative evaluation of the same qualitative research project. We utilized the research team's perspectives (data richness) regarding the interview data and examined its factor structure using the full scope of LIWC-22 variables. Largely motivated by Robinson et al. (2013), a three-step procedure for data analysis and factor analysis to predict the research team's rating score of the interview transcripts was applied. We explore the practicability of applying data analytics to identify subtle, hidden patterns based on linguistic features in interview transcripts that can be used to improve future data collection practices.

Through the use of exploratory factor analysis and multiple regression analysis, this study reveals specific linguistic patterns that are associated with the richness of data as perceived by the research team, pinpointing "Refined/Reflective Storytelling" and "Contextualized Relationships/Conflicts" as key factors influencing higher evaluation scores. The results demonstrate the utility of LIWC-22 as an insightful tool for enhancing interview techniques and

boosting the quality of qualitative research in projects. This investigation highlights the capability of data analytics tools such as LIWC-22 to support evaluators and researchers in more effectively and efficiently performing formative evaluations. It also sets the stage for further exploration into the use of automated text analysis for assessing qualitative data.

Significance of the Dissertation

The broader dissertation contributes to the body of RoE literature by providing three empirical studies that inform evaluation practice. This work answers previous calls for more RoE by a) building on previous efforts to categorize and learn from existing RoE and b) continuing the trend of identifying new tools to help evaluators be efficient and effective in their practice.

In the first study, the synthesis of RoE context provides empirical evidence on how evaluation context influences evaluation efforts and informs evaluation practices. Drawing off existing frameworks for evaluation context, we extend recent efforts to empirically examine the RoE literature more in depth. This study seeks to build evaluator competency as they respond to diverse contexts, engage with different stakeholders, explore the goals and political/historical aspects of the program and organization, and facilitate the evaluation process. With a more comprehensive understanding of the influence of evaluation context elements, evaluators will be better prepared to navigate a new context.

In the other two research studies, we draw data from a real-world evaluation of a qualitative research project to examine whether LIWC-22 is a viable evaluation tool. The second paper first shows how the software can be used to monitor and test for interviewer effects during data collection. Then, the third paper shows how evaluators can use LIWC-22 to identify variables that are associated with what is valued by the research team. In both cases, these studies show how evaluators can utilize LIWC-22 to define their formative evaluation plans for a

large qualitative research project. Identifying and demonstrating the use of such a tool for evaluation is an important first step in prescribing practice. The approach has the potential to decrease their cognitive bias and efficiently assist in the process of decision-making during data collection.

Importantly, although the focus of articles two and three is on evaluating research practices, qualitative data collected during large evaluations may just as easily be used to examine evaluation practices, particularly if variables aligned to important evaluation outcomes can be defined. Currently, LIWC has four summary variables that capture constructs that are important in psychological and social science research. Given the interaction between evaluators and stakeholders throughout the evaluation effort, the use of LIWC and existing summary variables may be important to inform qualitative interactions tied to context assessment, stakeholder selection, standard setting, and interpreting results. However, LIWC dictionaries are also customizable and, with future research in this vein, could be tailored to measure evaluation-specific constructs that may help practitioners with more nuanced evaluation activities like meta-evaluation, for example, through textual analysis of evaluation reports.

Methodologically, there is no existing research in the evaluation field that incorporates LIWC; however, several recent studies do exist that examine the use of tools like concept mapping (Trochim & McLinden, 2017); Geographic Information Systems (Azzam & Robinson, 2013); crowdsourcing (Azzam & Harman, 2016; Harman & Azzam, 2018), and data visualization (Evergreen & Metzner, 2013). This dissertation continues the trend of examining the use of tools to help evaluators improve their practice.

This dissertation is motivated by the need for more RoE. It presents three examples of RoE to advance the evaluation field further. The goal is to generate a detailed and comprehensive

guide for understanding the evaluation context and provide evaluators with a more systematic manual for proceeding when they start a new evaluation project. Much more work is needed, but the collective body of work provides important empirical evidence to help evaluators efficiently and systematically perform evaluation activities.

Limitations

The limitations of the study are as follows. For the first article, the literature that was utilized consisted of existing RoE articles that had already been categorized, and therefore is somewhat limited and dated, as it includes only the years 2005-2019. There is the possibility that researchers conducted RoE but did not publish it in any of the 14 journals covered by Coryn et al. (2017), for example, it may appear in book chapters, dissertations or other discipline-specific journals. Additionally, it is acknowledged that RoE findings can change based on the use of the studies and their dissemination. Not every evaluation article will include the necessary information to assess each dimension of context, but that does not mean it was not addressed or important in the broader evaluation effort. Also, the Vo (2013) framework of evaluation context guides the whole study, if using a different framework, there is a chance of getting different results.

For the second and third articles, the project dataset has a particular purpose, and not every evaluation requires high levels of authenticity or emotional tone in their data collection effort. Additionally, the cleaning of the dataset to promote analysis disregards possible sources of variability, such as interjections from interviewers. Furthermore, in some cases, related responses may have been addressed in questions outside of the three focal questions used for analysis. A more controlled environment could add consistency in the data collection process, which is important for detecting interviewer effects. Additionally, a different context, different

questions, or even different forms of text capturing (written vs. oral) may produce different results, considering the LIWC-22 output variables depend heavily on the length of the responses analyzed. Furthermore, this is especially true for the third article, as we asked the research team to rate the interview transcripts. Thus, the outcome variable has less construct validity given that what the research team values in terms of data richness may not align with what others value. In both of these cases, the use of LIWC is seen as a data analytics tool to help identify areas of possible concern which may require further investigation including ongoing monitoring, qualitative data analysis. The aim of adapting the preset LIWC package to support an evaluation effort should be viewed as supplemental to improve efficiency assist in decision-making, and not as a tool to provide definitive proof. Given the flexibility of the software to eventually expand to tailored dictionaries (e.g., dictionaries involving evaluative terms), the use of the preset LIWC, although a limitation, should be viewed as a minimal demonstration of its possible value, which can only increase with the development of refined measures.

CHAPTER 1 [ARTICLE 1]: THE ROLE OF CONTEXT: A SYNTHESIS OF EMPIRICAL RESEARCH ON EVALUATION CONTEXT

Zhi Li *Carl Westine*

Numerous evaluation scholars and theorists have noted the importance of investigating evaluation theories, methods, and practices for their possible benefits. This process, named "research on evaluation" (RoE), was defined as "any purposeful, systematic, empirical inquiry intended to test existing knowledge, contribute to existing knowledge, or generate new knowledge related to some aspect of evaluation processes or products, or evaluation theories, methods, or practices" (Coryn et al., 2016, p.161). According to Mark (2008), research on evaluation is the conduit for helping evaluators select and defend evaluation practice decisions from a multitude of options.

Starting from the 1970s, studies on evaluation began appearing in published literature (e.g., Bernstein & Freeman, 1975; Patton et al., 1977; Weiss, 1977). These studies primarily focused on research evaluation use. After a while, the momentum fizzled, and research on evaluation declined until the beginning of the 21st century (e.g., Alkin, 2003; Henry & Mark, 2003). However, since a call for more RoE by Christie (2003), it has regained scholars' and theorists' attention (Alkin, 2003; Coryn & Westine, 2015; Hansen et al., 2013; Mark, 2008; Vo, 2013). Since this call, the number of RoE studies has steadily grown over time (Coryn et al., 2017; Vallin et al., 2015; Webb et al., 2017). However, the evaluation field still lags well behind other social sciences in the pursuit of developing an evidence base to inform practice. As noted in their synthesis, evaluators need to also learn from the existing research on evaluation (Coryn et al., 2017).

This paper is an extension of Coryn et al. (2017). The former study aimed to classify research on evaluation using existing taxonomies from Henry and Mark (2003) and Mark (2008).

It fell short of capturing and evaluating what has been learned from research on evaluation and for each research on evaluation domains (i.e., evaluation context, evaluation activities, evaluation consequences, and professional issues). Given the importance of RoE to the field (Coryn et al., 2016; Szanyi et al., 2013), there is a need to dig deeper into each of the RoE domains to synthesize and define practice from empirical findings. The present study builds off Coryn et al. (2017) with a more intentional synthesis of the evaluation context domain. Through understanding the evaluation context, the researchers aim to find better ways to help evaluators use information that is relevant to the context and inform their decision-making about the evaluation process.

Among various evaluation scholars and theorists, Henry and Mark (2003) provided and later Mark (2008) refined an influential RoE taxonomy to assist in clarifying what is studied in research on evaluation, which includes evaluation context, evaluation activities, evaluation consequences, and professional issues. Later, Coryn et al. (2017) performed a systematic review of peer-reviewed RoE literature published in 14 evaluation-focused journals from 2005 to 2014 and tried to identify the proportion of the articles that fit these two taxonomies. The review highlighted several weaknesses in the earliest taxonomy with overlapping categories, making the more recent taxonomy more useful, though somewhat incomplete. Specifically, using Mark's (2008) subjects of inquiry taxonomy, which includes the domains of evaluation context, evaluation activities, evaluation consequences, and professional issues, Coryn et al. (2017) found broad coverage of RoE across the domains. According to the authors' results, more than half of the included RoE research articles belonged to evaluation activities (N=132, 51.36%), followed by evaluation consequences (N=70, 27.24%), evaluation context (N=58, 22.57%), professional issues (N=51, 19.84), and others (N=44, 16.34%). However, the review focused only on

categorizing RoE using the taxonomies and did not do any synthesizing of the literature findings within the taxonomies. There is a need to not only document and categorize the RoE literature but also to learn from its findings. Evaluation researchers should synthesize the studies found within the Mark (2008) subjects of inquiry domains to provide a clearer picture of what has been learned from the RoE during this time frame.

Context is relevant in all evaluations. Thus, research on evaluation context is an essential precursor to improve evaluation practice. This study concentrated on the evaluation context domain. Evaluation context was explained as the "circumstance within which evaluation occurs" (see Mark, 2008, TABLE 6.1). Given the tendency to conduct studies actively or retrospectively in most RoE context studies, the implication is that in most RoE contexts, the researchers examine the background and foundation within which an evaluation occurred or collectively a review of multiple contexts. Full awareness of the context can help build a comprehensive background of the evaluation, understand the role of the stakeholders and evaluators, choose and design the appropriate evaluation activities, and set up the proper evaluation standards and goals.

Review of Literature

Defining Evaluation Context

Based on the results of Coryn et al. (2017), most of the research on evaluation can be categorized as evaluation activities, and other areas, including evaluation context, are important in the field. Vo (2013) notes that in the evaluation field, scholars and theorists collectively agree that contextual factors are embedded in evaluation activities and associated scholarly efforts. Vo and Christie (2015) further explain that context plays a crucial role in shaping evaluation practice; however, until recently, there have been only a few frameworks to identify and understand the evaluation context. To address this knowledge gap, Greene (2005), Mark (2008), Hansen et al.

(2013), and Vo (2013) each presented their own versions of subcategories regarding evaluation context.

In the field of evaluation, Greene (2005) defines the context as "the setting within which the evaluand (the program, policy, or product being evaluated) and thus the evaluation are situated. Context is the site, location, environment, or milieu for a given evaluand" (p. 83). Fitzpatrick (2012) further points out that instead of a general understanding of the context of evaluation, some scholars tend to identify evaluation context within the cultural setting when it comes to evaluation practice. For example, Chouinard and Cousins (2009) define context as subsuming culture, noting that context is "the site of confluence where program, culture, and community connect" (p. 461).

Each of the authors noted above defines their categorizations of context differently, but similarities and refinements do exist. Greene (2005) identified five dimensions of context in evaluation, which included demographic characteristics of the setting and the people in it, material and economic features, institutional and organizational climate, interpersonal dimensions or typical means of interaction, and norms for relationships in the setting, and political dynamics of the setting, including issues and interests. There were three subcategories in Mark (2008): societal level, organizational level, and evaluation specific. Hansen et al. (2013) demonstrated four sections, including evaluator, organization/program, stakeholders, and others. Based on Greene (2005), Mark (2008), and Hansen et al. (2013), Vo (2013) generated her own evaluation context dimensions with the evaluator, stakeholder, organizational, program, and historical/political.

Naturally, all four scholars emphasize the significance and value of the organization in reference to the evaluation context. Evaluators and stakeholders also represent context domains

for the four scholars. However, according to Mark (2008), background characteristics of a specific evaluation are also a part of the evaluation context (p.118). This includes internal and external evaluators, evaluators' training and experiences, the history of evaluation in the local context, and the stakeholders' background. This explanation of evaluator and stakeholder is comparable to what Vo (2013) defines in her "evaluator descriptor" and "stakeholder descriptor" in the evaluation context dimension. That dimension includes an evaluator's methodological and interpersonal skills, content knowledge, values, and theoretical orientation that the evaluator possessed during their work, as well as a stakeholder's audience, identity, values, information needs, and expertise/skills.

There are also some subtle similarities in how the scholars classify evaluation context. Greene (2005) and Vo (2013) suggested a historical/political dimension in the evaluation context, including the historical events and the dynamics of relationships that have influenced and will continue to influence the development and direction of program evaluation. Carefully examining both of their explanations of the historical/political dimension, it is possible to see some correlation between what Mark (2008) termed societal level and evaluation-specific domains. Segerholm (2003) mentioned that evaluation could be conducted in a particular national context or, more generally, that evaluation could be performed at a societal level, for example, as "cross-national comparisons" (Mark, 2008, p. 118). The national, local, historical, and cultural contexts will shape and affect the evaluation process.

Even though all four scholars provided classifications for evaluation context, there has yet to be a synthesis of the details and dimensions of the evaluation context based on the peer-reviewed scholarly RoE articles. Thus, there is a need to understand the impact of specific

evaluation context descriptors on the evaluation process and its outcomes. Specifically, the study seeks to answer the following research question:

1. What does empirical research on evaluation context reveal about the influence of context in evaluation studies?

Theoretical Framework: Evaluation Context

As shown in Figure 1-1, the present study utilizes the framework closely aligned with Hansen et al. (2013) and Vo (2013) and characterizes the evaluation context dimensions of the evaluator, stakeholder, organizational, program, and historical/political.

Evaluator context encompasses all the skills and knowledge they need to apply in their work, which includes methodological and interpersonal skills, as well as expertise in specific content areas. It also embraces their values, their understanding of the evaluation's ultimate goal, and their theoretical approaches.

Vo (2013) refers to "stakeholder context" as the condition that when individuals work under or are influenced by the program, those involved in the evaluation might be recognized as the audience. Accordingly, "stakeholder context" is not only limited to the identities of the individuals included in the evaluation process but should also include detailed information about their information needs, values, and expertise. In the stakeholder context, the audience is a higher-level and integrated descriptor, including the content of other descriptors.

The term 'program context' refers to the characteristics of the program under evaluation. It details the size of the program, its developmental phase, and the human and material resources needed for its operation. Unlike stakeholders, programs are managed by individuals and possess distinct informational needs and values. Programs are established and operated based on specific objectives or missions that individuals aim to achieve through them. Here, simultaneously

considering organizational context and program context is reasonable because they mirror one another. The only difference between the organization and the program is that the organizational context is at a higher level than the program context. Ultimately, organizational values can assist in supporting, guiding, influencing, and setting up the program's mission.

Lastly, the "historical/political" dimension consists of "historical events" and "relationships." "Historical events" include policy initiatives that can give rise to programs. Relationships have influenced and will continue to influence the shape of the program under evaluation, falling under the political dimension. These relationships, whether between individuals or groups, can be internal to the program or represent links to entities outside of the program.

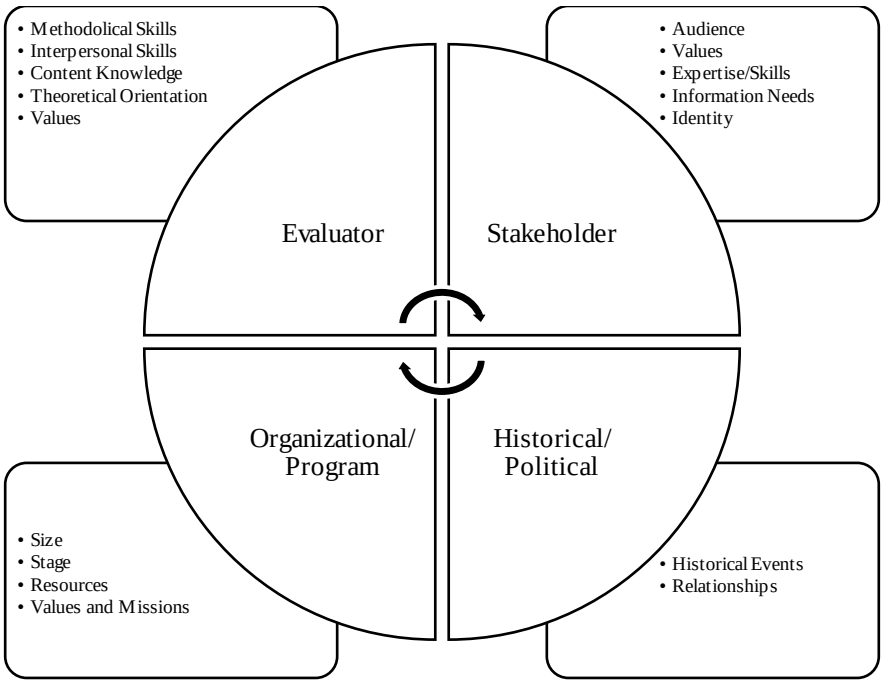

Figure 1-1. Framework for Evaluation Context (Source: Vo, 2013, p. 46).

Method

Sample

This study uses two sets of RoE context articles. The first set includes 58 peer-reviewed scholarly RoE context articles published between 2005 and 2014, identified by Coryn et al. (2017). These articles were categorized under the RoE context category following Mark's (2008) taxonomy. The second set consists of 14 peer-reviewed scholarly RoE context articles published from 2015 to 2019, identified by Linnell and Stachowski (2024).

Present practices to identify and categorize RoE studies are notoriously challenging and time-consuming. The methodology outlined by Coryn et al. (2017) involved a team first extracting all RoE articles from 14 evaluation-focused journals and then identifying which of them tied explicitly to the evaluation context domain. Repeating this process in its entirety was deemed infeasible for updating the dataset. However, current efforts to extend Coryn et al. (2017) up through 2019 using a more focused set of journals were actively occurring (e.g., Linnell & Stachowski, 2024; Prescher et al., 2023). Those ongoing efforts involved the classification of RoE articles from the *American Journal of Evaluation*, *Canadian Journal of Program Evaluation*, *Evaluation Review*, and *Journal of MultiDisciplinary Evaluation* which overlap with Coryn et al. (2017). Through collaboration we were able to get access to a list of an additional 14 published articles focusing on RoE context.

Thus, we examined the combined set of 72 articles to understand how each evaluation context dimension affects evaluation practice. Out of the list, 15 systematic reviews were excluded from this study because they did not contain information on how each context dimension affects evaluation practice; they only synthesized the evidence to their specific research questions. Additionally, not all the remaining 57 articles address each dimension of the

evaluation context, as Vo (2013) defined. This approach was mirrored in the analysis of the second set of papers. Thus, each dimension was considered separately and consisted of a different sample size. For example, in the case of Organization/Program, the most frequently addressed dimension, 23 additional papers were excluded because they did not adequately describe the organization or program. Figure 1-2 describes the number of articles considered with each evaluation context dimension, which ranges from 14 to 34.

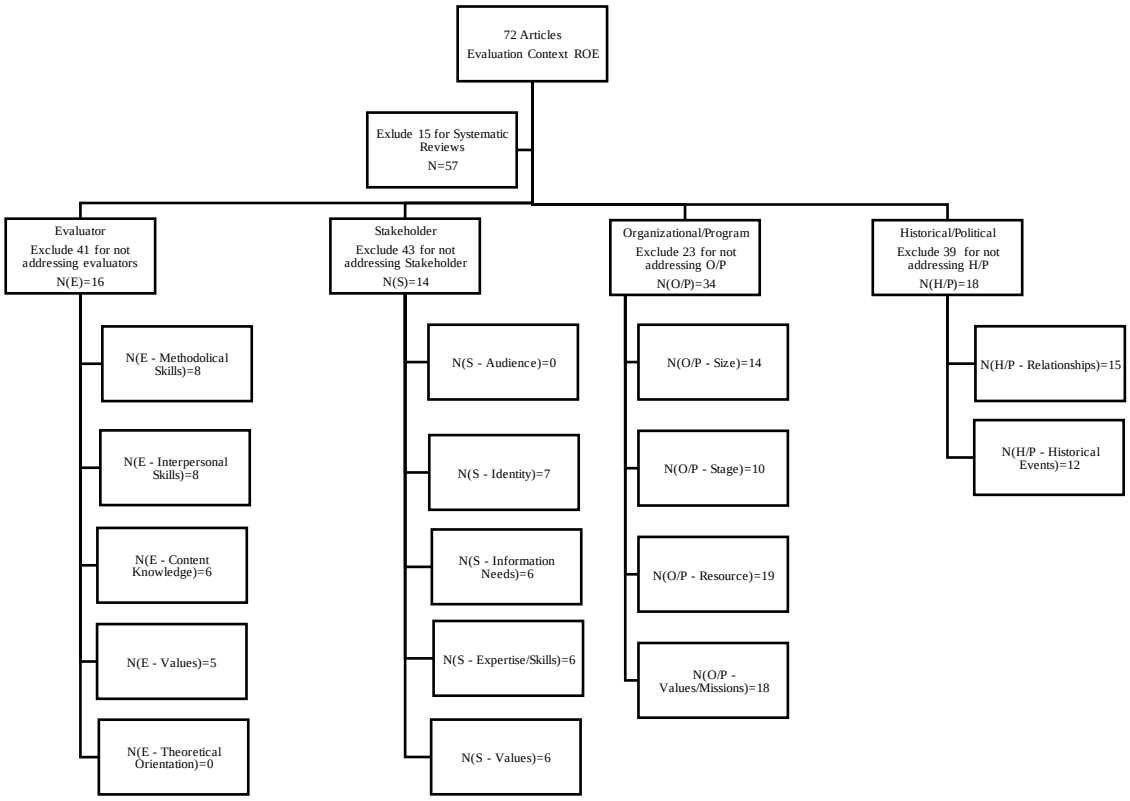

Figure 1-2. Sample Sizes for Context Dimension

Analysis Procedure

This study was carried out by a research team comprising a faculty member and a doctoral student. All articles were treated as qualitative data, guided by Lester et al. (2020). The analysis involved a deductive coding process guided by the framework established by Vo (2013).

Fifty-seven research articles were encoded using NVivo 1.7.1 (Lumivero, 2022) for detailed examination.

The coding was executed in two phases. The initial phase focused on identifying the dimensions of the evaluation context, as per Vo (2013), while the subsequent phase involved a more profound exploration within each dimension to categorize the descriptors. Specifically, paragraphs from the articles that concentrated on outcomes—such as results, discussion, and conclusions—were analyzed for content relevant to the various dimensions of the evaluation context. These dimensions encompass evaluator, stakeholder, organizational, program, historical, and political aspects.

Text segments within each dimension were then thematically grouped based on descriptors or their impacts on the evaluation process. Once the paragraphs were coded for each dimension and corresponding descriptors, an Excel spreadsheet facilitated cross-checking. The research team's constant discussion and collaboration enhanced the findings' trustworthiness throughout the analytical journey.

Findings

This section synthesizes the evaluation context dimensions within our selected framework adapted from Vo (2013), drawing upon an in-depth analysis of 57 articles. It aims to elucidate each contextual dimension and its descriptors, offering a comprehensive understanding of their implications for evaluation practices. Below is the table showing examples of quotes on how we code the text into each of the context descriptors (Table 1-1).

Table 1-1

Examples of Quoted Text's Assignment to Context Descriptors

Context Dimension	Context Descriptors	Example Quote
Evaluator	Methodological Skills	"All design choices should be made with care because they can influence the real and perceived validity and credibility of the entire evaluation. As a community we should be open to designing and implementing evaluations, regardless of our comfort level. Although it is not appropriate to use methods without properly understanding their strengths and limitations, this openness to methodological diversity can provide an opportunity for practicing evaluators to continue their professional growth and training" (Azzam, 2011, p. 389).
	Interpersonal Skills	"Communication skills are also key in enabling the evaluator to take a fuller role in project decision making. Evaluators who struggled with communication were not as readily invited to participate in decision-making processes, such as sustainability planning. They also were not frequently seen as full partners by the wider group of stakeholders" (Cartland et al., 2008, p. 476).
	Content Knowledge	"The advantages were being closer to and more familiar with the programs, being more familiar with the rhythms and cycles when evaluation activities might be embraced or resisted, developing a better understanding of the information needs of stakeholders, and dealing with fewer people and programs. The primary disadvantage was the potential of being less objective and introducing bias in evaluating programs" (Lambur, 2008, p. 51).
	Values	"Data from this study suggest that an evaluator who values and is well trained in experimental methods may be a better choice for this kind of evaluation work, rather than an evaluator with substantive expertise in education but no value for and by extension expertise in experimental methods" (Christie & Fierro, 2012, p. 71).
	Theoretical Orientation	None

Table 1-1 Examples of Quoted Text's Assignment to Context Descriptors (continued).

Context Dimension	Context Descriptors	Example Quote
Stakeholder	Identity	"Stakeholder role contributed 3–9% of the variance in assessments of LRC services, a small to medium effect. Post hoc comparisons indicated that teacher-librarians and curriculum support staff rated all LRC functions significantly higher than other groups; the differences were especially large between teacher-librarians and members of student support services (e.g., educational assistants). Multimedia design functions were given higher ratings by members of the Director's Office and curriculum support staff than other groups. Elementary teachers rated most LRC functions higher than secondary teachers (the exceptions were the virtual library and multimedia which were given equivalent ratings by both teacher groups)" (Ross, 2008, p. 360).
	Information Needs	"Where persistent/entrenched differences in perspective between stakeholders translate to different information needs (i.e., in terms of what is credible, their level of decision-making, and timing of decision-making cycles), incorporating multiple methods and tailored reporting into the evaluation design may help" (Svensson et al., 2018, p. 470).
	Expertise/Skills	"Professional Expertise: Outside professionals in specific categories can fill the roles in the government organization in the 7 categories above with the role of meeting professional standards in these roles and to advocate for more work on the basis of professional quality" (Lempert, 2010, Table 2, p. 70).
	Values	"Finally, some of the interviewees perceive the evaluation and how it is carried out as unfair, which triggers negative attitudes toward the evaluation, such as those disclosed in the following words: ''So now, some people sitting at their desks, who have no idea of what it is like to work with children, come to tell me on the basis of writing a portfolio that I can be rated as efficient, deficient, inadequate, or another rating that they came up with, that revolts me and gives me a stinging, because I find it totally unfair'' (Teacher No. 1)" (Tornero & Taut, 2010, p. 136).
	Audience	None

Table 1-1 Examples of Quoted Text's Assignment to Context Descriptors (continued).

Context Dimension	Context Descriptors	Example Quote
Organizational/Program	Size	"Finally, larger organizations, as measured by income, are more likely to use evaluation results to satisfy funder requirements and to obtain a seal" (Eckerd & Moulton, 2011, p. 112).
	Stage	"We can speculate that stable organizations are more ready to consider the kind of double-loop changes suggested by an external party. Another reason for our observation is presumably related to the scale of the evaluations that are usually outsourced. Because outsourcing is mainly reserved for large-scale evaluands, “solid” organizations will be better positioned to implement these. The involvement of an external evaluator is usually more expensive than conducting an internal evaluation" (Pattyn & Brans, 2013, p. 53).
	Resources	"The purposeful socialization into the organization’s evaluation process was mostly unsuccessful. While some limited professional development and coaching in evaluation was provided, no clear expectations for people’s evaluation roles were established, no tangible incentives for participation in the evaluation process were offered, and no “learning evaluation by doing it” was effectively promoted" (Volkov, 2008, p.192).
	Values and Missions	"Well-designed organizations have a clear mission and are structured in such a way as to focus energy toward achievement of that mission" (Lambur, 2008, p. 42).

Table 1-1 Examples of Quoted Text's Assignment to Context Descriptors (continued).

Context Dimension	Context Descriptors	Example Quote
Historical/Political	Historical Events	"Adding to the general challenges of implementing program evaluation in Indian country, participants expressed difficulty in collecting evaluation data due to the contextual factors of their programs. Specifically, participants noted the culture and history of their American Indian/Alaska Native communities and the nature and logistics of their Physical Activity Programs (n=6) as contextual factors that impacted data collection" (Roberts et al., 2018, p. 172).
	Relationships	"Maybe the main conclusion of our study is that in order to have an appraisal system that works according to the Western logics, you need to be in a condition of peace inside, as well as outside the organization. If an organization is under the constant threat of military actions, if the supplies are in constant danger of not arriving, if the employees experience problems of mobility on the territory, there is little chance that a performance appraisal system would motivate or control people" (Giangreco et al., 2012, p. 167).

Findings highlight specific examples that demonstrate and underscore various descriptors within each context dimension. Within the RoE articles examined (Figure 1-2), the most commonly identified dimension included the organizational/program context, represented in 60% of the articles. Noted examples demonstrate a varying influence from each dimension as to how evaluations are planned, executed, and received. Having more of any one quantitative descriptor (size, stage, or resources) or the presence of underlying values and missions does not consistently dictate decision-making with respect to defining evaluation scope or practice.

Other context dimensions appeared less frequently in the articles examined, but the identified examples do confirm the significance of each dimension. However, a few descriptors were not addressed in the sampled articles (Evaluator: Theoretical Orientation and Stakeholder:

Audience). Additionally, most of the descriptors were not as well-defined in practice as in the organizational/program dimension (i.e., easily interpreted directionally or as presence y/n). In some cases, authors' descriptor examples overlapped across dimensions (e.g., when addressing staff), signifying more specificity in classifying complex intersections of dimensions is needed.

Evaluator Context

Evaluators' skills and perspectives fundamentally shape the evaluation process and outcomes. This dimension encompasses methodological skills, interpersonal skills, content knowledge, values, and theoretical orientation, each playing a pivotal role in conducting effective evaluations.

Methodological Skills

Evaluators' methodological skills play a crucial role in the evaluation process. Their ability to design and implement appropriate research designs, select relevant evaluation methods, and analyze data accurately can greatly impact the quality and validity of the evaluation findings (Azzam, 2011). Methodological skills enable evaluators to ensure scientific rigor and maximize the internal and external validity of the evaluation (Whitesell et al. 2018). However, it is important to consider the context in which the evaluation is being conducted and the adaptability of the chosen methods to that specific setting (Chouinard & Hopson, 2015). Evaluators should also be aware of the potential unintended consequences of their methodological choices, such as excluding relevant findings or compromising cultural rigor (Cheng & King, 2017). While the employment of stringent research methodologies, like randomized controlled trials, is frequently emphasized in evaluations, it's crucial to examine if these methods are universally the most suitable for assessing all types of interventions (Tornero & Taut, 2010). Overall, evaluators'

methodological skills are essential for producing credible and reliable evidence in the evaluation process.

Interpersonal Skills

The interpersonal skills of evaluators play a significant role in the evaluation process. Building personal relationships and rapport with colleagues is crucial for successful evaluation practice (Cartland et al., 2008; Vanderkruik & McPherson, 2017). By leveraging interpersonal dynamics at work, internal evaluators can enhance the evaluation skills of team members and boost the organization's capacity for evaluation, as suggested by Rogers et al. (2019). Evaluators should interact with individuals on their own terms, guiding discussions and encouraging introspection without taking over, as highlighted by Cheng & King (2017). Ensuring all voices are heard, not just the most dominant ones, leads to a more inclusive and effective evaluation method, according to Pattyn & Brans (2013). The skill of presenting evaluation outcomes in the organization's vernacular increases the likelihood of those findings being used, as Pattyn (2014) notes. External evaluators, too, can convey information effectively by engaging with stakeholders in a collaborative manner. Overall, evaluators' interpersonal skills contribute to building trust, fostering collaboration, and increasing the likelihood of evaluation utilization and impact.

Content Knowledge

Evaluators' content knowledge significantly impacts the evaluation process. Having expertise in the subject matter being evaluated allows evaluators to understand the nuances and complexities of the program or intervention being assessed (Lambur, 2008). This knowledge enables evaluators to ask relevant and insightful questions, design appropriate evaluation methods, and interpret findings accurately (Carman & Fredericks, 2008). It also helps evaluators

to identify potential biases and limitations in the evaluation process and address them effectively. Additionally, evaluators with content knowledge can provide valuable insights and recommendations based on their understanding of the field, contributing to the overall quality and usefulness of the evaluation (Cheng & King, 2017; Pattyn & Brans, 2013; Rogers et al., 2019; Shaw & Faulkner, 2006).

Values

The values held by evaluators can have an impact on the evaluation process. Evaluators' values can influence the criteria and standards used to assess the effectiveness of a program or intervention. They can also shape the interpretation of evaluation findings and the recommendations made based on those findings (Christie &Fierro, 2012). Evaluators' values can affect the emphasis placed on different aspects of the evaluation, such as the importance given to stakeholder perspectives or the focus on quantitative data versus qualitative data. Additionally, evaluators' values can influence how evaluation results are communicated and used by stakeholders. For example, evaluators strongly committed to social justice may prioritize equity and fairness in their evaluation practices. On the other hand, evaluators with a more technocratic orientation may prioritize efficiency and cost-effectiveness (Cheng & King, 2017; Kallemeyn et al., 2015; Vanderkruik & McPherson, 2017). There is no information specifically tied to the theoretical orientation descriptor.

Stakeholder Context

The stakeholder context in an evaluation process includes many elements that affect how an evaluation is seen, done, and used. It is shaped by the stakeholders' identities, information needs, expertise, skills, and values.

Identity

Stakeholders' identity can have an impact on the evaluation process. Different stakeholder groups may have different perspectives, interests, and needs, which can influence their evaluation of a program or initiative. For example, teachers may focus on the needs of service providers, while administrators may prioritize the organization's needs as a whole (Svensson et al., 2018). The position of the evaluator within the system can also affect their perspective and potential biases (Bundi, 2016). Additionally, stakeholders' power dynamics and relationships can influence their involvement and participation in the evaluation process (Ross, 2008). Evaluators need to recognize and consider these different stakeholder perspectives and identities in order to ensure a comprehensive and inclusive evaluation (Kovač & Langfeldt, 2010).

Information Needs

Stakeholders' information needs can have a significant impact on the design and implementation of evaluations. Understanding the specific information needs of stakeholders is crucial for ensuring that the evaluation process is relevant and useful to them. By incorporating multiple methods and tailored reporting into the evaluation design, evaluators can address the diverse information needs of stakeholders (Svensson et al., 2018). Additionally, stakeholder analysis can help identify and prioritize these needs, allowing for more targeted and effective evaluation efforts (Pleger et al., 2017). Allocating resources to coaching stakeholders throughout the evaluation process can also help ensure that their information needs are met (Bundi, 2016). Furthermore, evaluation can play a role in clarifying the focus of accountability, helping stakeholders understand the purpose and potential outcomes of the evaluation (Kovač, & Langfeldt, 2010). By considering stakeholders' information needs, evaluations can be designed to

provide the necessary information and insights to support decision-making and improve program effectiveness (Greenseid & Lawrenz, 2011).

Expertise/Skills

Stakeholders' expertise and skills can significantly influence the evaluation process. Their involvement in the evaluation can bring valuable insights and perspectives that contribute to a more comprehensive and accurate assessment of the program or intervention being evaluated (Lempert, 2010). By actively participating in the evaluation, stakeholders can provide firsthand knowledge and experiences that can inform the evaluation findings and recommendations (Labin, 2014). Additionally, stakeholders' expertise in specific areas related to the intervention can enhance the evaluation's technical expertise and ensure that the evaluation is conducted using the most appropriate methods and approaches (Kovač & Langfeldt, 2010). Additionally, the expertise of stakeholders in analyzing data, interpreting results, and making informed decisions plays a pivotal role in the successful application of evaluation outcomes and the execution of suggested actions (Greenseid & Lawrenz, 2011). Overall, stakeholders' expertise and skills play a crucial role in shaping the evaluation process and outcomes, making their involvement essential for a comprehensive and meaningful evaluation (Fleischer& Christie, 2009)

Values

Stakeholders' values can have a significant impact on the evaluation process. When stakeholders have different values, it can lead to conflicts and disagreements in interpreting evaluation findings and making decisions based on them. These differences in values can influence the level of acceptance or rejection of evaluation conclusions (Tornero & Taut, 2010). Stakeholders might dismiss the outcomes of evaluations due to personal beliefs and values, instead of analyzing the evidence (Svensson et al., 2018). Additionally, stakeholders' values can

shape their priorities and preferences in the evaluation process. Involving stakeholders in the evaluation can help balance political agendas and ensure that their values are taken into account (Ross, 2008). However, it is important to note that stakeholders as decision-makers may not always base their decisions directly on evaluation information, as their values and other factors may also play a role in decision-making (Kovač & Langfeldt, 2010). Overall, stakeholders' values can significantly influence the evaluation process and the use of evaluation findings. There is no information linked to the descriptor of audience.

Organizational/Program Context

The organizational/program context is pivotal in shaping an evaluation's approach, scope, and outcomes. Descriptors, as the organization's size, stage of development, available resources, and underlying values and missions, directly influence how evaluations are planned, executed, and received.

Size

Organizational/program size can have an influence on evaluation. Big organizations, with their ample resources and greater acceptance of potential risks, may see an enhancement in their evaluation capabilities and outcomes (Chouinard & Hopson, 2015; Eckerd & Moulton, 2011). Conversely, smaller entities, known for their agility, reduced bureaucracy, and openness to innovation and change, might also experience an increase in evaluation capacity and effectiveness (Bourgeois et al., 2016). The location or placement of the evaluation function within the organizational hierarchy can also impact evaluation. Centralized evaluation structures may have a few staff members responsible for evaluation, while decentralized structures may involve program staff in evaluation activities. Hybrid structures that combine centralized technical support with decentralized program staff involvement can successfully ensure

stakeholder involvement and produce quality evaluations (Ross, 2008). Overall, the relationship between organizational/program size and evaluation is complex and requires further study to understand the specific conditions contributing to evaluation capacity and performance (Hsu & Hsueh, 2009; Lambur, 2008).

Stage

The stage of an organizational/program's development can have an impact on the evaluation process. In the early stages, there may be a lack of awareness or understanding of the importance of evaluation, leading to limited resources and attention being given to evaluation efforts. As the organization/program progresses, there may be a greater recognition of the need for evaluation and a shift towards more systematic and comprehensive evaluation practices. This can include the establishment of evaluation advisory committees, integrating evaluation into daily work routines, and monitoring evaluation efforts (Pattyn & Brans, 2013; O'Connor & Netting, 2008). Additionally, the stage of an organization/program can also influence the level of stakeholder involvement in the evaluation process, with early stages often characterized by limited participation and later stages involving more diverse and inclusive stakeholder engagement (Lu et al., 2019)

Resources

Organizational/program resources play a crucial role in evaluation. The availability of resources, such as funding, personnel, and technology, can impact the design and implementation of evaluation activities. Organizations with limited resources may struggle to conduct comprehensive evaluations or may need to prioritize certain aspects of evaluation over others (Rogers et al., 2019; Volkov, 2008). Additionally, the level of administrative support for evaluation within an organization can influence the use of evaluation information in decision-

making processes (Ross, 2008). The location or placement of the evaluation function within the organizational hierarchy can also affect its credibility and importance (Lambur, 2008). Furthermore, the context of a program, including factors such as the presence of a crisis or conflict, can shape the usage, sense-making, and results of performance appraisal systems (Hedler & Gibram, 2009). Overall, the availability and support of resources within an organization can significantly impact the evaluation process and outcomes.

Values and Missions

The values and missions of an organization or program can significantly impact the evaluation process. These values and missions shape the goals and objectives of the organization, which in turn influence the criteria and standards used to assess the effectiveness and impact of programs and policies (Lambur, 2008). When the values and missions prioritize accountability, transparency, and evidence-based decision-making, evaluations are more likely to be rigorous, comprehensive, and unbiased. On the other hand, if the values and missions prioritize other factors, such as political considerations or maintaining the status quo, evaluations may be limited in scope, biased, or even disregarded altogether. Therefore, it is crucial for organizations and programs to align their values and missions with a commitment to evaluation as a tool for learning, improvement, and accountability (Rogers et al., 2019; Tarsilla, 2014; Westbrook et al., 2017).

Historical/Political Context

Historical Events

Historical events can have a significant impact on the field of evaluation. In some countries, evaluation practices face challenges due to historical and cultural factors, which might lead to unpopularity or skepticism towards evaluation methods (Roberts et al., 2018).

Additionally, evaluation societies have evolved over time, with different societies emphasizing value-neutral processes, complexity, and quantifiable information for decision-making (Kallemeyn et al., 2015). The increase in impact evaluation can be traced back to the historical development of performance audits and the demand for evidence-based policy and research (Tornero & Taut, 2010). Evaluations have also become an important instrument to assess public policies, with parliamentarians demanding evaluations to hold the government accountable (Ledermann, 2012). Resistance to evaluation can arise from factors related to the dispositions of those being evaluated, situational aspects, and the evaluation itself. Overall, historical events shape the context, perception, and utilization of evaluation in various settings.

Relationships

Political relationships can have a significant impact on evaluation practices. The context in which evaluations take place, including the political environment, can shape the emphasis on use and valuing in evaluation (Kallemeyn et al., 2015). Evaluators are not simply responding to contextual factors but are actively constructing and engaging with the context in which they operate (Bundi, 2016). Evaluations are often demanded by parliaments as a means of providing information for decision-making and fulfilling their oversight function (Boehmer & Zaytsev, 2019). The accountability of programs to various stakeholders, including political leaders, influences the legitimacy and effectiveness of evaluations (O'Connor & Netting, 2008). In some cases, evaluations may be used as a tool for holding the government accountable and controlling bureaucratic agencies (Giangreco et al., 2010). The political leadership style and centralization of power can also influence the evaluation process and outcomes. Overall, political relationships play a crucial role in shaping the demand for evaluations, the focus of evaluations, and the use of evaluation findings.

Discussion

This study embarked on a comprehensive empirical exploration of contemporary RoE with a focused lens on evaluation context. By synthesizing findings across the evaluator, stakeholder, organizational/program, and historical/political dimension of contexts, we have extended the foundational work of Coryn et al. (2017) and engaged with broader thematic discussions using frameworks guided by Vo (2013).

Our study aligns with the critical need Coryn et al. (2017) identified for a more nuanced understanding of RoE domains. However, our approach diverges by documenting and offering a deeper analysis of how specific context descriptors—evaluators' methodological skills, interpersonal skills, content knowledge, values; the stakeholders' identities, skills, information needs, and values; organizational/program size, stage, resources, value; and historical events/political relationships—affect the evaluation process. This examination allows us to contribute innovative insights to the RoE literature, emphasizing the complexity and the multifaceted nature of the context. Additionally, value is a recurring theme in Vo (2013)'s Evaluation Context framework, placing it at the core of the notion that evaluation fundamentally involves assessing value. This emphasis aligns consistently with definitions of evaluation as the determination of something's merit and worth, as described by Scriven (1991) and the American Evaluation Association (2014).

A critical distinction between our work and Vo (2013), Greene (2005), and others lies in the empirical nature of work, drawing off specific RoE examples. While Vo (2013) and Greene (2005) provided valuable frameworks for understanding evaluation context, our study delves deeper into each context dimension, employing a qualitative analysis to a broad set of recent RoE studies. This approach reaffirms the significance of contextual factors identified in previous

studies which present frameworks of context, but also in certain instances give additional detail as to how these factors can influence evaluation practices.

One important finding is that many of the dimensions of evaluation context are not adequately represented in the published RoE articles. Within the 57 articles considered, the descriptors within the organizational/program dimension were most frequently described, but still absent in a majority of articles. In many of the other context dimensions attention to descriptors was quite limited if not absent in some cases. While it is likely that a research article format naturally limits description of specific domains like the evaluator and stakeholders (Vo & Christie, 2015), this information is still vital to the practice and advancement of RoE.

To further meta-level research like the present study which draws off existing primary studies, evaluators need consistency in reporting certain information that can later be used to for synthesis purposes and examining the influence of key factors and dimensions. Recent efforts to promote reporting standards for evaluation (Montrosse-Moorhead & Griffith, 2017) and even constructing cases for the teaching of evaluation (Linfield & Tovey, 2021; Kallemeyn et al., 2021; Tovey & Greene, 2021) need also to be applied to RoE to promote standardizing dissemination and maximizing utility among secondary researchers and RoE users.

Our findings do underscore the importance of methodological and interpersonal skills, echoing the sentiments of Whitesell et al. (2018) and Rogers et al. (2019), while also highlighting the critical role of evaluators' content knowledge and values, a descriptor underexplored in earlier RoE literature. Furthermore, by examining the stakeholder context, we illuminate how stakeholder identities, information needs, expertise, and values collectively shape the evaluation landscape, a perspective that adds depth to the discussions initiated by Svensson et al. (2018) and Ross (2008).

In the organizational/program context, our analysis extends the dialogue on the influence of size, stage, resources, and missions/values on evaluation, providing empirical evidence to complement theoretical discussions by Chouinard and Hopson (2015) and Lambur (2008). Lastly, our exploration of the historical/political context adds to the narrative by showcasing how historical events and political relationships frame the evaluation environment, building on insights from Roberts et al. (2018) and Bundi (2016).

Hence, our study's key takeaways also emphasize evaluation contexts' dynamic and interdependent nature. We illustrate that while individual context dimensions have a unique influence, their collective interplay can also shape evaluation outcomes. This holistic view encourages evaluators and scholars to consider the broader evaluative system, ensuring that evaluations are methodologically sound, culturally sensitive, stakeholder-inclusive, and organizationally relevant.

Conclusions

The findings from this study offer significant contributions to the understanding of the four dimensions in the evaluation context as depicted within RoE studies. They highlight the critical need for standardizing how context is described and disseminated in publications. They also emphasize the complex interplay between the evaluator, stakeholder, organizational/program, and historical events/political relationships.

For the evaluator dimension, the study underscores the critical role of evaluators' methodological skills, interpersonal abilities, content knowledge, and values in shaping the evaluation process. These elements collectively influence the credibility, inclusiveness, and effectiveness of evaluations. Evaluators' capacity to adapt methodologies to the context of each project, coupled with their interpersonal skills in engaging diverse stakeholder groups, enhances

the utility and impact of evaluation findings. Moreover, the alignment of evaluators' values with ethical and equitable evaluation practices further ensures that evaluations contribute constructively to program improvement and stakeholder understanding.

For the stakeholder dimension, our research highlights the importance of acknowledging and integrating stakeholders' varied identities, information needs, expertise, and values throughout the evaluation process. Acknowledging these varied viewpoints not only enhances the design and execution of the evaluation but also promotes an evaluation setting that is more inclusive and responsive. Stakeholders' active participation and the alignment of evaluation processes with their needs and expectations are essential for enhancing the relevance and utilization of evaluation outcomes.

For the organizational/program context dimension, findings indicate that the size, stage of development, resource availability, and the values and missions of the organization or program exert a significant influence on the evaluation process. While the size and available resources provide a foundational capacity for conducting evaluations, it is the stage of development and the alignment with organizational values and missions that determine the extent to which evaluations can effectively inform decision-making, policy development, and program improvement efforts.

For the historical/political dimension, the study sheds light on the profound impact of historical events and political relationships on evaluation practice. Historical experiences, particularly among marginalized communities, and the prevailing political climate influence both the approach to and reception of evaluation efforts. Recognizing and addressing these historical and political factors is crucial for conducting evaluations that are not only methodologically sound but also culturally sensitive and politically aware.

Implications

The findings from this study present some practical suggestions for both researchers and evaluators. Specifically, they highlight that context can be viewed as overarching to an evaluation effort (e.g., for interpreting its conclusions), or as relative to the evaluator or program (e.g., in the planning stages) for decision-making purpose. For example, when the evaluation team enters a new evaluation project, it needs to understand the key aspects of this project's context, including detail associated with the organization and program, stakeholders, and relevant the historical and political factors. Understanding this information enables evaluators to better position and coordinate around their overall fit, enabling them adapt to the context they face. This would better equip them to initially navigate the evaluation situation they are in and inform their decision-making about selecting the best evaluation approach for the specific evaluation effort or add team members with requisite experience.

Likewise, a program should carefully consider these dimensions in searching for and selecting an evaluator. From the program's perspective, there is a fundamental need to bring in the evaluator that best fits with the evaluation context, as this could otherwise represent a significant threat to the overall utility of the evaluation. Programs also need to understand and anticipate how their contextual situation (size, stage, resources, values, and mission), stakeholders (audience, identity, information needs, expertise/skills, values), and historical/political context (historical events, relationships) will impact the evaluation effort to inform their selection of the right evaluator.

Two theoretical examples are presented for illustration purposes. First, consider the situation presented in Figure 1-3 where evaluative practice is valued and rooted in the organization/program's context, and the program has a duty to respond to many specific

stakeholder information needs. In this case it might be viewed as more important to have an evaluator with a theoretical orientation aligned to collaborative and participatory evaluation approach which prioritizes evaluation capacity building yet responsive to the needs of stakeholders with sustained involvement in practices such as information-gathering, analysis, and interpretation. Moreover, in this case, the presence of any historical/political factors, although not specifically identified in the example may refine or influence how much collaboration the evaluation chooses to include.

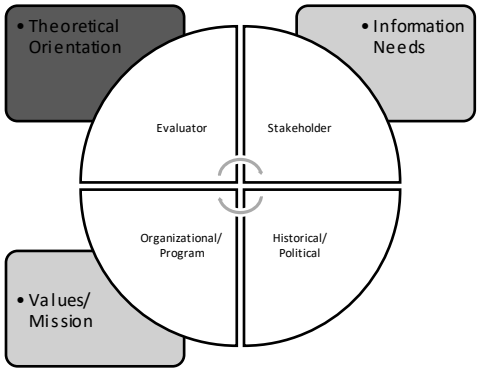

Figure 1-3. Example 1

In another scenario (Figure 1-4), consider an evaluation context that involves a lot of historical and political relationships within a relatively large program. In this situation it may be beneficial to have evaluator with exceptional interpersonal skills to navigate the communication complexities. The program administrators may seek to prioritize this type of skill when identifying the ideal evaluator.

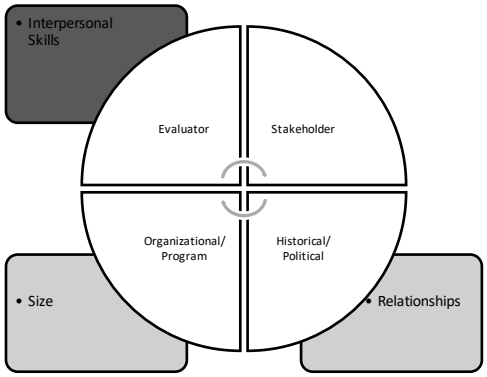

Figure 1-4. Example 2

When conducting an evaluation, evaluators and program administrator should pay attention to the specific features of the organization and the program, such as size, age, staff, budgets, mission, and other characteristics. As noted above, even the smaller elements of context could be considered when hiring an evaluator or making the evaluation plan as it may affect the feasibility or outcome of the evaluation. Evaluators conducting an evaluation should intentionally adapt their designs to better account for contextual dimensions and descriptors.

Limitations

The research articles we looked at include a mix of evaluation cases, specific studies, and collections of studies within particular contexts, which presents a challenge to aggregate and generalize. More information about each dimension and descriptor in the articles is needed to understand the influence of context in evaluation studies. Moreover, there are also many overlaps in the categories. For example, staff can be human resources, the number of staff can indicate the organization and program's size and stage, staff can be both evaluator and stakeholder in evaluation, and there is value descriptor across evaluator, stakeholder, and organizational

dimension. So, there is a need for more definitions of the categories to guide the study process and a shared understanding of context descriptors for each dimension.

There is also a need to define and measure critical intersections of specific context dimensions rather than focusing only on measuring individual descriptors, as many overlaps exist among the descriptors. Moreover, the scope of this study is limited to the years 2005-2019. It is possible that researchers conducted RoE studies but did not publish their findings in the 14 journals analyzed by Coryn et al. (2017). Instead, these studies may have appeared in alternative formats like book chapters or dissertations. Although this time frame is significant because it aligns with the resurgence of RoE, additional empirical RoE studies have been produced over the years since 2014, and ongoing efforts to identify and classify RoE continue. It is important to acknowledge that RoE findings could change based on the use of more studies derived through different definitions of RoE or studies from different time periods. Efforts to expand RoE should continue to distill findings through syntheses addressing what is learned from the RoE to replicate these findings.

References

American Evaluation Association. (2014). What is evaluation? https://www.eval.org/Portals/0/What%20is%20evaluation%20Document.pdf

Alkin, M. C. (2003). Evaluation theory and practice: Insights and new directions. In C. A. Christie (Ed.), *The practice–theory relationship: New directions for evaluation* (Vol. 97, pp. 89–91). San Francisco, CA: Jossey-Bass.

Bernstein, I. N., & Freeman, H. E. (1975). *Academic and entrepreneurial research*. New York: Russell Sage.

Boehmer, H. M., & Zaytsev, Y. K. K. (2019). Raising aid efficiency with international development aid monitoring and evaluation Systems. *Journal of MultiDisciplinary Evaluation, 15*(32), 28-40.

Bourgeois, I., & Cousins, J. B. (2013). Understanding dimensions of organizational evaluation capacity. *American Journal of Evaluation, 34*, 299–319.

Bundi, P. (2016). What do we know about the demand for evaluation? Insights from the parliamentary arena. *The American Journal of Evaluation*, *37*(4), 522–541. https://doi.org/10.1177/1098214015621788

Cheng, S.-H., & King, J. A. (2017). Exploring organizational evaluation capacity and evaluation capacity building: A delphi study of Taiwanese elementary and junior high Schools. *The American Journal of Evaluation*, *38*(4), 521–539. https://doi.org/10.1177/1098214016672344

Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation, 30*, 457–494.

Chouinard, J. A., & Hopson, R. (2015). A critical exploration of culture in international

development evaluation. *Canadian Journal of Program Evaluation*, *30*(3), 248-276. doi: 10.3138/cjpe.30.3.02

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice–theory relationship: New directions for evaluation* (Vol. 97, pp.7–36). San Francisco, CA: Jossey-Bass.

Coryn, C. L. S., Ozeki, S., Wilson, L. N., Greenman, II, G. D., Schro¨ter, D. C., Hobson, K. A., . . . Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation, 37,* 159–173.

Coryn, C. L. S., & Westine, C. D. (Eds.). (2015). *Contemporary trends in evaluation research* (Vols. I-IV). (Sage benchmarks in social research methods). London, England: Sage.

Coryn, C. L. S., Wilson, L. N., Westine, C. D., Hobson, K. A., Ozeki, S., Fiekowsky, E. L., … Schröter, D. C. (2017). A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation*, *38*(3), 329–347. doi: 10.1177/1098214016688556

Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. In D. J. Rog, J. L. Fitzpatrick, & R. F. Conner (Eds.), *Context: A framework for its influence on evaluation practice. New Directions for Evaluation*, *135*, 7–24.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation*, *30*(2), 158–175. https://doi.org/10.1177/1098214008331009

Giangreco, A., Carugati, A., Sebastiano, A., & Tamimi, H. A. (2010). War outside, ceasefire inside: An analysis of the performance appraisal system of a public hospital in a zone of

conflict. *Evaluation and Program Planning*, *35*(1), 161–170. https://doi.org/10.1016/j.evalprogplan.2010.11.004

Greene, J. C. (2005). Context. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 82–84). Thousand Oaks, CA: Sage.

Greenseid, L. O., & Lawrenz, F. (2011). Using citation analysis methods to assess the influence of science, technology, engineering, and mathematics education evaluations. *American Journal of Evaluation*, *32*(3), 392-407.

Hansen, M., Alkin, M., & Wallace, T. (2013). Depicting the logic of three evaluation theories. *Evaluation and Program Planning*, *38*(C), 34–43. https://doi.org/10.1016/j.evalprogplan.2012.03.012

Hedler, H. C., & Ribeiro, N. G. (2009). The contribution of metaevaluation to program evaluation: Proposition of a model. *Journal of MultiDisciplinary Evaluation, 6*(12), 210-223.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.*), The practice–theory relationship in evaluation: New directions for evaluation* (Vol. 97, pp. 69–80). San Francisco, CA: Jossey-Bass.

Hsu, F.-M., & Hsueh, C.-C. (2009). Measuring relative efficiency of government-sponsored R&D projects: A three-stage approach. *Evaluation and Program Planning, 32*, 178-186. doi:10.1016/j.evalprogplan.2008.10.005

Kallemeyn, L. M., Hall, J., Friche, N., & McReynolds, C. (2015). Cross-continental reflections on evaluation practice: Methods, use, and valuing. *The American Journal of Evaluation*, *36*(3), 339–357. https://doi.org/10.1177/1098214015576400

Kallemeyn, L. M., Titiml, E. N., Galib, L., Castelin, K., Montrosse-Moorhead, B., Ensminger, D. C., & Linfield, K. J. (2021). What are the characteristics of the cases we use to teach and learn evaluation? *New Directions for Evaluation*, *2021*(172), 19–35. https://doi.org/10.1002/ev.20479

Kleiman, S. (2004). Phenomenology: To wonder and search for meanings. *Nurse Researcher*, *11*(4), 7–19. doi: 10.7748/nr2004.07.11.4.7.c6211

Kovač, V. B., & Langfeldt, G. (2010). Educational evaluation in the light of construal level theory: The case of cognitive tuning. *Studies in Educational Evaluation*, *36*(3), 93–100. https://doi.org/10.1016/j.stueduc.2011.01.001

Labin, S. N. (2014). Developing common measures in evaluation capacity building: An iterative science and practice process. *The American Journal of Evaluation*, *35*(1), 107–115. https://doi.org/10.1177/1098214013499965

Lambur, M. T. (2008). Organizational structures that support internal program evaluation. 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*, 41–54.

Lempert, D. (2009). Why government and non-governmental policies and projects fail despite 'evaluations': An indicator to measure whether evaluation systems incorporate the rules of good governance. *Journal of MultiDisciplinary Evaluation,6*(13), 58-108.

Lester, J. N., Cho, Y., & Lochmiller, C. R. (2020). Learning to do qualitative data analysis: A starting point. *Human Resource Development Review*, *19*(1), 94–106. https://doi.org/10.1177/1534484320903890

Linfield, K. J., & Tovey, T. L. (2021). What is the case for teaching with cases in evaluation? *New Directions for Evaluation*, *2021*(172), 11–18. https://doi.org/10.1002/ev.20480

Linnell, D. J. & Stachowski, A. (2024). *The next eight-years of published research on evaluation: A follow-up of Coryn et al. (2017)* [Unpublished manuscript].

Lu, S. K., Elliot, S. J., & Perlman, C. M. (2019). Perceived facilitators and barriers to evaluative thinking in a small, development NGO. *Canadian Journal of Program Evaluation*, *34*(1), 68-83. https://doi.org/10.3138/CJPE.43118

Lumivero (2022). *NVivo* (1.7.1). www.lumivero.com

Mark, M. M. (2008). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111–134). New York, NY: Guilford.

Montrosse-Moorhead, B., & Griffith, J. C. (2017). Toward the development of reporting standards for evaluations. *American Journal of Evaluation, 38*(4), 577–602. https://doi.org/10.1177/1098214017699275

O'Connor, M. K., & Netting, F. E. (2008). Faith-based evaluation: Accountable to whom, for what? *Evaluation and Program Planning*, *31*(4), 347–355. https://doi.org/10.1016/j.evalprogplan.2008.04.013

Patton, M. Q., Grimes, P. S., Guthrie, K. M., Brennan, N. J., French, B. D., & Blyth, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in public policy making* (pp.141–184). Lexington, MA: Lexington Books

Pattyn, V., & Brans, M. (2013). Outsource versus in-house? An identification of organizational conditions influencing the choice for internal or external evaluators. *28*(2), 43-63.

Pattyn, V. (2014). Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation, Canadian Journal of Program Evaluation*, *20*(3), 348–367. https://doi.org/10.1177/1356389014540564

Pleger, L., Sager, F., Morris, M., Meyer, W., & Stockmann, R. (2017). Are some countries more prone to pressure evaluators than others? Comparing findings from the United States, United Kingdom, Germany, and Switzerland. *The American Journal of Evaluation*, *38*(3), 315–328. https://doi.org/10.1177/1098214016662907

Prescher, M., Sulze, K., Linnell, D., & Stachowski, A. (2023, October 9-14). *The story of research on evaluation (RoE): Continuing a thorough investigation on recent trends* [Poster presentation]. Evaluation 2023, Indianapolis, United States.

Roberts, E. B., Butler, J., & Green, K. M. (2018). Challenges to evaluating physical activity programs in American Indian/Alaska Native communities. *The American Journal of Evaluation*, *39*(2), 166–182. https://doi.org/10.1177/1098214017733544

Rogers, A., Kelly, L. M., & McCoy, A. (2019). Evaluation literacy: Perspectives of internal evaluators in non-government organizations. *The Canadian Journal of Program Evaluation*, *34*(1), 1-20. https://doi.org/10.3138/CJPE.42190

Ross, J. A. (2008). Cost–utility analysis in educational needs assessment. *Evaluation and Program Planning*, *31*(4), 356–367. https://doi.org/10.1016/j.evalprogplan.2008.06.003

Scriven, M. (1991). *Evaluation thesaurus*. Sage

Segerholm, C. (2003). Researching evaluation in national (state) politics and administration: A critical approach. *American Journal of Evaluation*, *24*, 353-372.

Shaw, I., & Faulkner, A. (2006). Practitioner evaluation at work. *The American Journal of Evaluation*, *27*(1), 44–63. https://doi.org/10.1177/1098214005284968

Svensson, K., Szijarto, B., Milley, P., & Cousins, J. B. (2018). Evaluating social innovations: Implications for evaluation design. *The American Journal of Evaluation*, *39*(4), 459–477. https://doi.org/10.1177/1098214018763553

Szanyi, M., Azzam, T., & Galen, M. (2012). Research on evaluation: A needs assessment. *Canadian Journal of Program Evaluation*, *27*(1), 39–64. https://doi.org/10.3138/cjpe.027.002

Tarsilla, M. (2014). Evaluation capacity development in Africa: Current landscape of international partners' initiatives, lessons learned and the way forward. *African Evaluation Journal*, *2*(1), e1–e13. https://doi.org/10.4102/aej.v2i1.89

Tornero, B., & Taut, S. (2010). A mandatory, high-stakes national teacher evaluation system: Perceptions and attributions of teachers who actively refuse to participate. *Studies in Educational Evaluation*, *36*(4), 132–142. https://doi.org/10.1016/j.stueduc.2011.02.002

Tovey, T. L., & Greene, J. C. (2021). What is next for cases in teaching and learning evaluation? A call to action. *New Directions for Evaluation*, *2021*(172), 103–108. https://doi.org/10.1002/ev.20485

Vallin, L. M., Philippoff, J., Pierce, S., & Brandon, P. R. (2015). Research-on-evaluation articles published in the *American Journal of Evaluation*, 1998–2014. In P. R. Brandon (Ed.), *Research on evaluation*. *New directions for evaluation* (Vol. 148, pp. 7–15). San Francisco, CA: Jossey-Bass

Vanderkruik, R., & McPherson, M. E. (2017). A contextual factors framework to inform implementation and evaluation of public health initiatives. *The American Journal of Evaluation*, *38*(3), 348–359. https://doi.org/10.1177/1098214016670029

Vo, A. T. (2013). Visualizing context through theory deconstruction: A content analysis of three

bodies of evaluation theory literature. *Evaluation and Program Planning, 38*, 44-52. doi:10.1016/j.evalprogplan.2012.03.013

Vo, A. T., & Christie, C. A. (2015). Advancing research on evaluation through the study of context. In Paul R. Brandon (Ed.), *Research on evaluation. New Directions for Evaluation, 148,* 43–55.

Webb, A. L., Schumacker, R. E., & Tilford, A. (2017). Synthesis of published articles from studies in educational evaluation, 2010-2015. *Journal of Education and Human Development*, *6*(1), 78–81. doi: 10.15640/jehd.v6n1a7

Weiss, C. H. (Ed.). (1977). *Using social research in public policy making*. Lexington, MA: Lexington Books.

Westbrook, T. R., Avellar, S. A., & Seftor, N. (2017). Reviewing the reviews: Examining similarities and differences between federally funded evidence reviews. *Evaluation Review, 41*(3), 183–211. https://doi.org/10.1177/0193841X16666463

Whitesell, N. R., Sarche, M., Keane, E., Mousseau, A. C., & Kaufman, C. E. (2018). Advancing scientific methods in community and cultural context to promote health equity: Lessons from intervention outcomes research with American Indian and Alaska Native communities. *The American Journal of Evaluation*, *39*(1), 42–57. https://doi.org/10.1177/1098214017726872

CHAPTER 2 [ARTICLE II]: USING DATA ANALYTICS TO MONITOR AND EVALUATE QUALITATIVE DATA COLLECTION PROCESSES FOR INTERVIEWER EFFECTS

Zhi Li *Carl Westine*

Grant-funded projects frequently require external evaluation as a condition of funding. High-quality evaluation, in theory, will positively affect the equity and efficaciousness of the project results. Today, it is also commonplace to have research efforts attached to grant-funded projects, necessitating the evaluator to not only track and evaluate project initiatives and objectives but also to monitor the progress of research efforts and assess the quality of the research products. In particular, when the research component of a grant is prominent, initial efforts to monitor and report on data collection could be viewed as a fundamental step in establishing evaluator credibility with the client, which is essential for promoting future evaluation activities and leading to better evaluation use. In this study, the authors test one strategy that evaluators can use to evaluate ongoing research efforts formatively. Using a grant-funded project with a qualitative research aim, we examine the feasibility of using data analytics to report on data collection involving a large number of successive interviews by a small team of interviewers.

An evaluator typically plays various roles throughout the evaluation effort (Skolits et al., 2009; Volkov, 2011), but a central role during the active evaluation phase involves collecting and analyzing data aligned to merit criteria to judge the evaluand against standards (Scriven, 1993; Fournier, 1994). Formally, evaluators determine the merit, worth, or significance of an evaluand (Scriven, 1982).

The evaluand for a grant is defined by the scope of the project and related research components. Grant writers will typically articulate project objectives, the procedures required to

fulfill these aims, including any intended research efforts, as well as a request for the required resources to fund the project. Project evaluators frequently track the implementation and outcomes of these efforts to improve project performance and judge the success of project elements. In evaluating the research element of a project, similar attention to the research processes and outcomes is also needed. Importantly, an evaluation of a grant project typically constitutes a fraction of the overall grant budget, implying the funding for evaluation is at a cost to the projects' bottom line. Thus, evaluators are not only required to provide sound practices but also be efficient in their resource use (Yarbrough et al., 2010). In particular, when the research component of a grant is prominent, initial efforts to monitor and report on data collection could be viewed as a fundamental step in establishing evaluator credibility with the client, which is essential for promoting future evaluation activities and leading to better evaluation use.

Good evaluation practice requires unbiased context-sensitive behavior (Alkin & Vo, 2018). However, coming from different backgrounds, evaluators bring their various experiences, knowledge, expertise, culture, and ethnicity into the evaluation project and apply appropriate methods and theories to design and conduct evaluation. Regrettably, within the common setting, both internal and external evaluators may exhibit a bias towards positive outcomes (Scriven, 1993). In a grant context, the relationship between the grantee and evaluator is particularly susceptible, given the source of funding for evaluation. Additionally, economic, social, political, and psychological pressures can push evaluators to be too positive, and the fear of having to report on negative outcomes (Davidson, 2015) can enable evaluator biases to dictate the direction of inquiry.

For years, the evaluation community has been calling for more research on evaluation (Christie, 2003; Henry & Mark, 2003; Mark, 2008; Coryn et al., 2016). Research on evaluation is

"any purposeful, systematic, empirical inquiry intended to test knowledge, contribute to existing knowledge, or generate new knowledge related to some aspect of evaluation processes or products, or evaluation theories, methods, or practices" (Coryn et al., 2016). These calls repeatedly point to a need for practice-based research to identify efficient and credible strategies and tools that can routinely be employed by evaluators in the field (Coryn et al., 2017). Formal examination of various strategies will generate evidence to inform and prescribe evaluation practice in particular contexts, thereby reducing the influence of evaluator biases. This, in turn, is an essential step toward the professionalization of the discipline. Research on evaluation helps to establish and improve the evaluation project while also assisting in creating a better research setting.

Skolits et al. (2019) state that credible and quality data is needed to address evaluation questions and justify evaluation decision-making competently. Hence, conducting a professional evaluation necessitates that the evaluator employs a methodical strategy to mitigate possible mistakes and guarantee the systematic gathering and analysis of data (Alkin & Vo, 2018). Thus, project evaluators need sound, flexible, transparent, and systematic strategies to formatively and summatively judge research practices. In the context of evaluating research efforts tied to a grant project, on one extreme, such an evaluation could be conceptualized as an independent reproduction of all of the research steps drawing off the full scope of the chosen evaluation model (e.g., assessing the need for the research to a complete reanalysis of the project data.) However, this practice is likely to be resource-intensive and not conducive to providing timely and formative information. Alternatively, to the other extreme, evaluation could be based on anecdotal reflections on data collection practices and products from the research team and study participants. But this limited view is subject to bias and prone to generating superficial evidence.

Unfortunately, such efforts are commonly employed given the accessibility of project staff, participants, and other stakeholders. A middle ground is needed. Evaluation requires evaluators to be free to pursue and investigate what is valued, particularly by key stakeholders, but systematic and efficient procedures that provide structure and practicality into their inquiry process. Through research on evaluation, evaluators can develop the necessary procedures and tools to assist research teams in improving their efforts and adding credibility to research findings.

In this study, the evaluand is a National Science Foundation (NSF) grant-funded qualitative research project that focused on "Studying Successful Doctoral Students in Mathematics from Underrepresented Groups (SSS)," and the tool is Linguistic Inquiry and Word Count (LIWC).

Review of Literature

Linguistic Inquiry and Word Count Software

Pennebaker and his colleagues initially created a text analysis tool named Linguistic Inquiry and Word Count (LIWC, pronounced "Luke") with the aim of improving the analysis of spoken and written language samples. This initial version of LIWC emerged from a research project exploring language use and disclosure, as detailed in Francis & Pennebaker (1992), but has evolved through several versions to establish dictionaries that enable researchers to efficiently examine qualitative data using only the counts of specific words. The simplistic use of word counts in defined LIWC dictionaries presented as a percentage of the total text length generates a consistent measure (Moore et al., 2019). As with previous versions, the most recent version, LIWC-22 (Pennebaker et al., 2022), is crafted for the swift and effective analysis of single or multiple language files. It is important to note that LIWC is not a tool to replace

qualitative analysis; instead, it utilizes quantitative data analytics to objectively identify important undertones in the text defined by word choices.

In addition to the more than 100 standard LIWC dimensions, which are based on the percentage of total words, four summary variables are available: analytical thinking, clout authenticity, and emotional tone. Each summary variable builds upon previously published research from Pennebaker's research team and therefore has established concurrent validity evidence (Humphreys & Wang, 2018). The software's unique algorithm automatically computes the measures, which are then transformed into percentiles using standardized scores from Pennebaker's comprehensive database for comparison (Boyd et al., 2022). In line with earlier editions, the summary variables remain the sole opaque elements in the LIWC-22 output. Despite being recalibrated according to new standards, these summary variables maintain conceptual continuity with the scores derived in LIWC2015.

Two specific summary variables are of interest in the present study: Authenticity and Emotional Tone. Each has been described and used in several recent research studies (e.g., by Oliver et al., 2020). For the present study, each variable is conceptualized as an essential indicator of high-quality data collection.

Authenticity (M. L. Newman et al., 2003). According to Oliver et al. (2020), the authenticity variable in LIWC can help to identify inauthenticity in the author's message. Interviewees changing their message during a response are more likely to receive a score closer to zero in authenticity. Whereas interviewees who are "personal, humble, and vulnerable" are likely to receive scores closer to 100 (Oliver et al. 2020, p. 336). The algorithm used to determine authenticity was developed from studies of honesty and deception, where it was found that participants who were lying were more likely to use "fewer self-references" and "more

negative words" (Oliver et al., 2020, p. 336). However, based on the LIWC Analysis website, as time goes by, the research team has come to a sense that "Authenticity measure has less to do with "deception" in a traditional sense and is, instead, more a reflection of the degree to which a person is self-monitoring." For example, lower scores in authenticity would suggest reading prepared texts and being cautious socially. At the same time, people who are more into spontaneous conversations with fewer social restrictions tend to get a higher authenticity score. In alignment with the research goal of the grant project, higher authentic scores are assumed to be an indicator of improved rapport between interviewer and interviewee, and therefore better interview data as the interviewee presumably feels more comfortable sharing their story with the interviewer.

Emotional Tone (Cohn et al., 2004). According to Oliver et al. (2020), the emotional tone variable in LIWC can help to identify positive and negative emotions within an author's message. While the algorithm is non-transparent, more positive emotion words are associated with a higher score, and more negative emotion words are associated with a lower score. This study assumes that a higher emotional tone score also indicates a closer connection between interviewer and interviewee has been established, leading to better interview data.

LIWC has been utilized in various fields, including Linguistics (Carroll, 2007; Imahori, 2018), psychology (Cutler et al., 2020; del Pilar Salas-Za´rate et al., 2014; Pennebaker & Francis, 1996; Pennebaker & Lay, 2002); Educational Technology (Crossley & McNamara, 2013; Geng et al., 2020). However, although several researchers have employed different versions of LIWC in other disciplines, no one has applied LIWC-22 in the evaluation context.

Conceptual Framework: Stufflebeam's CIPP Model

For this study, we draw upon the view that evaluation should consider more than just outcomes. Stufflebeam's Context, Input, Process, and Product (CIPP) evaluation model is one of the most widely known evaluation models (Esgaiar & Foster, 2019). It was created in the late 1960s to improve U.S. public school projects' accountability. Since then, the model has been further developed and applied across various disciplines and fields worldwide (Stufflebeam & Coryn, 2014). The CIPP Model is defined as "a comprehensive framework for conducting formative and summative evaluations of projects, personnel, products, organizations, and evaluation systems" (Stufflebeam & Shinkfield, 2007, p. 325). Unlike traditional evaluation, "learning by doing" roots in the CIPP model, which is described as "an ongoing effort to identify and correct mistakes made in evaluation practice, to invent and test needed new procedures, and to retain and incorporate especially effective practices" (Stufflebeam & Coryn, 2014, p. 310).

Zhang et al. (2011) describe the CIPP model as a series of evaluative inquiries that progressively guide the evaluator: What actions are required?; What is the best method to undertake these actions?; Are the actions being implemented?; and Has the initiative been successful? In the CIPP model, a context evaluation should focus on the program's needs, problems, and goals, with a primary emphasis on assessing the program's general fit in context (Al-Shanawani, 2019). Stufflebeam and Coryn (2014) explained what role evaluators, decision-makers, oversight bodies, and program stakeholders play in context evaluations. Evaluators assess the requirements, issues, resources, and related contextual circumstances. Decision-makers use context evaluation to set the goal and make sure the goal is defined to address the program's needs and problems. Oversight bodies and stakeholders use context evaluation to criticize if the appropriate goals guide the program and the outcome is responsive to the targeted

needs, problems, and goals. An input evaluation assists decision-makers in addressing the program's needs and meeting the program's implementation and operational goals. According to Stufflebeam and Coryn (2014), evaluators identify and assess the adequacy of resources to meet these needs, achieve implementation goals, and judge the program. This provides the necessary information to help decision-makers make plans regarding funding and allocating various resources and improve program planning. Process evaluation is used to monitor and judge the implementation of the program. Through evaluators' efforts in monitoring, documenting, giving ongoing feedback, and reporting on the program implementation process, program stakeholders learn the areas which need improvement and subsequently adjust their plans accordingly for better delivery (Stufflebeam & Coryn, 2014). Ultimately, during a product evaluation, the evaluator examines and gauges both the anticipated and unforeseen short-term and long-term accomplishments and results of the program. The evaluator offers suggestions of whether it is worth adopting the program, what elements need to be modified, and what improvement plans need to be made. At the end of the program, product evaluation assists in measuring the accomplishment, merit, worth, and significance of the program.

In this focused study, the evaluand is a NSF qualitative research project which documenting the experiences, perspectives, and stories of successful doctoral students and recent PhDs from historically marginalized racial groups in mathematics, including Black students, Latinx students, and Indigenous students. Specifically, we focus on aspects of the research data collection effort, which are articulated in Table 2-1 using the dimensions of the CIPP model (Stufflebeam & Zhang, 2017). First, context evaluation includes defining the goals and assessing contextual elements. For example, the term "context" refers to the conditions present for the interviews, which were conducted at various times, using "Zoom," during the Covid-19

pandemic, with many highly publicized racial injustices. Second, input evaluation involves evaluators' focus on the required elements, in this case, for data collection. Here, "input" refers to the demographic information of the specific interviewers and interviewees as well as the four interview protocols used for interviews which were aligned to the different stages of the math graduate program. Third, process evaluation should document, monitor, assess, and report on implementation. In the present study, "process" concerns the length of the interviews, paring of the interviewer and interviewee (e.g., by race or gender), and the use of the interview protocol, such as the specific order of the questions when they were asked during the interview, and the feedback provided to the interviewees by the interviewers. Fourth, product evaluation aims at the project's outcome and outputs, including judging the project's continuity and impact. In our narrow study, the set of interview data transcripts is viewed as the product which is affected by the other three dimensions. To ensure high quality, we can focus on evaluating various aspects of Context, Inputs, and Processes. In this study, we consider the presence of interviewer effects which are aligned to the "Input" and "Process" dimensions, as possible opportunities for improvement.

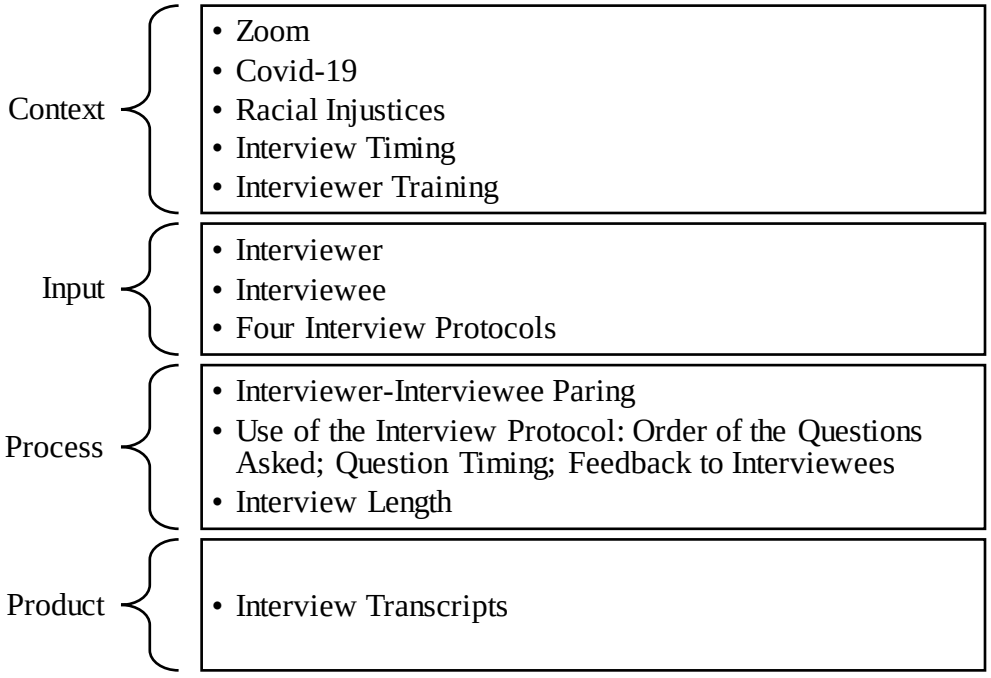

Figure 2-1. Factors Influencing CIPP Elements of the Grant-funded Project's Research Data Collection

Qualitative Data Collection/Quality (Interview)

This study is focused on evaluating a grant-funded qualitative research study on successful doctoral students' experience in mathematics from underrepresented groups. Inequities in education have prompted the need for more qualitative research into the lived experiences of minoritized and underserved students in Science, Technology, Engineering, and Mathematics (STEM) pipelines and administrators of STEM programs. Large research projects, such as those funded by the National Science Foundation, seek to document practices of doctoral programs from all angles to understand the context and program factors that facilitate the continual growth of minoritized graduates from STEM programs.

A key role of these efforts involves collecting and analyzing high-quality data that is full of in-depth descriptions. Given that the project objectives are built on developing a rich dataset,

project evaluators are obligated to assess the research processes in place to gather the qualitative data and evaluate the dataset as it is developed.

There are three major qualitative data resources: interviews, observations, and existing documents (Patton, 2015). Among these three common forms of qualitative data collection, interviews are the most prevailing way of collecting qualitative data (Merriam & Tisdell, 2016). In most qualitative research studies, "some and occasionally all of the data are collected through interviews," and "the most common form of interview is person to person encounter" (p.108). This is particularly true of equity studies, where critical race theory and related interview-based frameworks that place supreme value in the voice of minority students are used to understand the lived experience of these students. Hence, it is incumbent upon the research team to produce exceptional interviews.

Trustworthiness of Qualitative Research

To a large extent, high-quality qualitative research depends on complex interviewer-respondent interaction. The trustworthiness of the research is predicated on the processes in place to generate the dataset, which acknowledges that the biases and predispositions of both the interviewer and interviewee can "affect the interaction and the data elicited" (Merriam & Tisdell, 2016, p.130). Lincoln and Guba (1985) used the tenants of "credibility, transferability, dependability, and confirmability" to establish the trustworthiness of the overall research project; however, certain tenants are more applicable to a study on research processes. For example, with the credibility of qualitative research, Patton (2015) specifically points out that "the trustworthiness of the data is tied directly to the trustworthiness of those who collect and analyze the data—and their demonstrated competence" (p. 706). Researchers' practice, knowledge, and intelligence largely contribute to the credibility of qualitative research.

Good interview questions yield rich, thick, descriptive data; the more detailed, the better (Merriam & Tisdell, 2016). Centering good questions at the heart of the matter, the authors suggest that a researcher must ask good questions to collect meaningful data and stories about the phenomenon. Relatedly, Langley and Meziani (2020) mentioned that when an interviewer knows enough about the topic, they can easily detect what is going on, step back from the interview, and ask meaningful questions. Thus, the research process relies on good questions and the researcher's ability to understand the response in context, which helps the interviewer assess the quality of the response provided and rephrase or dig deeper as necessary.

Skilled navigation on the interviewer's part of which questions to ask and when to ask them is part of the interviewer-interviewee interaction and goes a long way toward bringing about a positive interaction and valuable response. In every interview situation, three variables determine the interviewer-interviewee interaction's nature: "(1) the personality and skill of the interviewer, (2) the attitudes and orientation of the interviewee, and (3) the definition of both (and often by significant others) of the situation" (Dexter, 1970, p. 24). Thus, one crucial role for the evaluation team is to perform critical analysis to generate reflection regarding the researcher's relationship to the study, any of the researchers' relevant assumptions or biases, the researchers' worldview and theoretical orientation, the nature of the interviewer-interviewee interaction, and the triangulation of the evidence being generated (Merriam and Tisdell, 2016, p. 259, TABLE 9.2).

However, specific strategies for conducting these analyses are not well-researched in an evaluation context. One possible option proposed in the present study includes an analysis of the nature of the data collection process (e.g., the depth of responses, time spent to achieve saturation, and presence of interviewer effects) concerning relevant demographic variables that

have been found to impact data collection like age, gender, ethnicity, and socioeconomic status of interviewers and interviewees (Foster, 1994). Seidman (2013, p.101) also found that power issues can affect the interview relationship, "who controls the direction of the interview, who controls the results, who benefits." Moreover, the sense of power is generally affected by "our experience with issues of class, race, and gender."

Interviewer Effects

As noted above, interviewers can and should be considered extensions of the interview protocol, impacting the social science data collection process. Existing studies have shown that interviewers can positively and negatively contribute to the quality of the data (Kühne, 2020). Interviewers have a significant influence during data collection, for example, dealing with complex instruments, clarifying interview questions, and guiding and motivating interviewees to answer accurately and provide rich answers (Loosveldt, 2008).

However, proper management of interviewers and their training can reduce the influence during data collection due to behaviors, interactions, and presence, often called "interviewer effects," which can harm the credibility and trustworthiness of the data collected (Kühne, 2020). An interviewer effect refers to the influence of an interviewer's characteristics like age, gender, race, skills, belief, and level of experience upon the responses provided by an interviewee (Leone et al., 2021) and has been prominently analyzed in the existing literature. Sensitive topics involving race/ethnicity (e.g., Adida et al., 2016; Kim et al., 2018; van Bochove et al., 2015), gender (e.g., Benstead, 2013; Fuch, 2009; Martin, 2020), sexual-related-behavior (Leone et al., 2021) are particularly prone to interviewer effects. The interviewer's workload (Wuyts & Loosveldt, 2020) and the educational gap between the interviewer and interviewee (Yang & Yu, 2008) can also be critical.

West and Blom (2017) conducted a meta-analysis on interviewer effects, and results showed that interviewers' race/ethnicity, age, gender, overall experience, and sociodemographic matching with respondent's characteristics are the top five characteristics that affect response quality (Table 4, p. 187). Results showed that to decrease measurement error, many studies suggest matching interviewer and interviewee by their race, but more work needs to be done when looking at the interviewer race effect as the question itself, the workload of the interviewer, and other mediators can affect interviewee. Additionally, the authors found an interviewer effect for gender. Their results showed that females tend to collect higher quality data as defined by hand-coding of responses for various desired textual properties.

Research Purpose

Given the ongoing push from the research on evaluation community to identify tools that can effectively and efficiently aid evaluators, in this study, we consider one option to assist in a current evaluation scenario: monitoring and evaluating qualitative data collection on a grant-funded, research-focused project. The presence of interviewer effects can have a detrimental impact on qualitative research projects. Thus, project evaluators should test for relevant interviewer effects in evaluating project-related research efforts. Early detection of the interviewer effects could minimize their impact and give the grant team options to improve their data collection practices.

This study proposes LIWC as one option for evaluators to test for interviewer effects. LIWC has pre-established summary variables that carry concurrent validity with factors indicative of high-quality qualitative data, making it an efficient choice. Additionally, LIWC and its existing measures provide the necessary structure and objectivity, limiting possible evaluator biases.

The purpose of the study is to explore the use of LIWC to assist evaluators in monitoring the data collection process and identifying high-quality data when evaluating a research project. Specifically, we utilized LIWC-22 to test interviewer effects, which can be important criteria in an evaluation with a focus on data collection and tracking of the research process. The following research questions guided this study:

1. To what extent do the authenticity and emotional tone scores of interviewee responses vary by question?
2. To what extent do the authenticity and emotional tone scores of interviewee responses vary by interviewer race and gender?
3. To what extent do the authenticity and emotional tone scores of interviewee responses vary by the alignment of interviewer/interviewee demographics (race, gender)?
4. Are there any interaction effects between question and interviewer demographics (race, gender) or the alignment of interviewer/interviewee demographics (race, gender) on authenticity and emotional tone score of interviewee responses?

Method

Research Design

This study used a correlational design to test for the presence of certain interviewer effects using interviewee responses to specific interview questions. The two dependent variables are LIWC-derived authenticity and emotional tone score for interviewee responses. The authors tested for differences in the authenticity and emotional tone scores of interviewee responses by question, demographics (race and gender) of the interviewer, and alignment of interviewer and interviewee demographics (race and gender). Specifically, the independent variables are the question asked by the interviewer (three levels: meaning of success; role that race played in math

experience; mentoring relationship), the interviewer's race (two levels: black versus other racial groups), the interviewer's gender (two levels: male versus female), the alignment of interviewer/interviewee race (two levels: same versus different), the alignment of interviewer/interviewee gender (two levels: same versus different).

The study focused on testing for the presence of the interviewer's main effects as well as the specific interactions between questions and demographic variables for each outcome. Thus, separate analyses derived from two-way ANOVAs were used for each outcome variable. In accordance with the two-way ANOVAs, five parallel hypotheses were formulated below, as shown in Table 2-1. Thus, a total of 9 hypotheses were tested for each outcome. Given the study's exploratory and formative nature, we left the significance threshold at alpha=.05 to identify possible opportunities for improvement but acknowledge the increased likelihood of significance given multiple tests.

Table 2-1

Hypotheses

Null hypothesis (H_0)	Alternate hypothesis (H_a)
• There is no significant difference in Authenticity/Emotional Tone score for each Question.	• There is a significant difference in Authenticity/Emotional tone score for each Question.
• There is no significant difference in the Authenticity/Emotional Tone score for Interviewer Demographic (Race/Gender).	• There is a significant difference in the Authenticity/Emotional Tone score for Interviewer Demographic (Race/Gender).
• There is no significant difference in Authenticity/ Emotional Tone score for interviewer/interviewee Demographic (Race/Gender) alignment.	• There is a significant difference in Authenticity/Emotional tone score for interviewer/interviewee Demographic (Race/Gender) alignment.
• There is no significant interaction between Question and Interviewer Demographic (Race/Gender) on Authenticity/Emotional tone score.	• There is a significant interaction between Question and Interviewer Demographic (Race/Gender) on Authenticity/Emotional tone score.
• There is no significant interaction between the Question and the interviewer/interviewee Demographic (Race/Gender) Alignment on Authenticity/Emotional tone score.	• There is a significant interaction between the Question and the interviewer/interviewee Demographic (Race/Gender) Alignment on Authenticity/Emotional tone score.

Dataset

For the purpose of evaluation in an organization and program context, we highlighted efforts on an NSF research project documenting the experiences, perspectives, and stories of successful doctoral students and recent PhDs from historically marginalized racial groups in

mathematics, including Black students, Latinx students, and Indigenous students. Historically, minoritized students have been highly underrepresented in mathematics doctoral programs and continue to face enormous challenges in their academic pursuits (American Mathematical Society, 2018; Okahana et al., 2018). The grant project research team collected 57 in-depth interviews using a diverse set of six interviewers. The interview team consisted of mathematics and educational researchers with diverse backgrounds across gender, race, discipline, and age. Among all the interviewers, two were male and four were female. Additionally, three identify as Black, while the other three interviewers identify as Asian, Latina and Multiracial. There were no white interviewers. Five of the interviewers have previous experience as math graduate students, and one interviewer has no math background. Each interviewer conducted a varying number of interviews (Table 2-2).

Table 2-2

Interviewers' Demographic Information

Interviewer	Race/Ethnicity	Gender	Number of Interviews
A	Black	Male	9
B	Asian	Female	13
C	Latina	Female	7
D	Black	Female	12
E	Black	Male	8
F	Multiracial	Female	8

Interviewee race/ethnicity was coded as Black/African American (including Caribbean) n=31, Latino/a (n=19), and multiracial (n=7). However, given the smaller sample size for some

subgroups, the interviewer race variable was recoded as a binary variable by combining the Asian, Latina and multiracial groups into one group.

All interviews were semi-structured and followed a general interview protocol aligned to four key topical areas: background, environment, mentoring, and academics. Each interview was recorded and transcribed and lasted between 60 and 200 minutes. After transcription, the interview transcripts were scrubbed for identifying information and assigned a random identifier. Within those transcripts, there were four cohorts of participants: newly accepted students, early graduate students (pre-qualifying exams), advanced graduate students (dissertation level), and recent PhDs (0-5 years since graduation). Depending on the cohort of the participants, a few of the interview questions were different based on the designated interview protocol. Moreover, due to the semi-structured interview format, each interviewer may have addressed the interview questions slightly differently based on the specific circumstances during the interview. Some questions were asked in a more standard way, while others were addressed more casually because of the nature of a conversation. Here are the examples for asking about "success":

1. Yeah. Yeah. That makes sense. How do you, what does success in graduate school mean to you? Or how do you like, I guess how were you envisioning that and what you think of on it?

2. I know that [inaudible 00:07:30]. You mentioned this word success, I think at least two times. And I just want to clarify, what does success mean to you? What does success mean to you an undergrad first?

3. Awesome. This is more like a [inaudible] question, but what does success in grad school mean to you?

The interviewer and interviewee were partnered randomly based on their available timeslots, which resulted in the fact that they frequently had different genders or were in different ethnic groups. Other possible differences existed among the pairings as well, such as being from diverse academic backgrounds or in different age groups, but these appeared less frequently.

Sample

Three questions were asked commonly across all the interviews: (1) What does success in graduate school mean to you? (2) How do you think race has played a role in your math experience? (3) What is the ideal mentoring relationship looks at to you? For the 57 interview transcripts, the responses were filtered to consider only cases where at least one of these three questions was asked in the conventional way and answered. Within all the transcripts, the desired information was available for 51 cases involving the success question, 46 for the race question, and 42 for the mentor question.

Analysis Procedure

Three major steps define the study process: (1) extracting and grouping the responses from the 57 transcripts based on responses to three commonly asked questions throughout the interviews, the interviewers' demographic information, and the pairing of interviewer and interviewee, (2) formatting the text responses into complete responses for input into LIWC-22 to check the score for Authenticity and Emotional Tone, and (3) statistically analyzing the data to investigate the interrelationship among variables.

Data cleaning and processing for interviewees showed 50.9% male representation and 42.1% Black representation in the entire dataset. Minor interjections in the transcripts, such as "Mm-hmm" or "That's great" were ignored and not included for the purposes of analysis. In these

cases, the interviewee's responses were combined to appear as one statement. Other cases like when the interviewer made a short acknowledgment statement to reflect what the interviewee shared, were also not included, and an interviewee's response was formatted to appear as one statement.

After importing the textual data into LIWC-22, a list of scores to be used as outcomes were generated using the LIWC-22 software. The analysis for research questions included two procedures: (1) A descriptive analysis of data to describe the distribution and test the assumptions, including homogeneity of variance, independence of observations, and the presence of a normally distributed dependent variable. (2) A two-way ANOVA hypothesis testing.

Results

In this section, results of the quantitative analysis are presented sequentially for the two LIWC outcome variables considered: authenticity and emotional tone. As such, readers are reminded that the analyses only utilize the converted scores of the transcribed interview responses and are, therefore, not a substitute for proper qualitative data analysis. Each analysis is only intended to inform formative evaluative practice with respect to data collection processes, and not as a measure of the quality of the collected data, which was graciously provided.

Outcome: Authenticity

Using the interviewee's authenticity score as the outcome, separate two-way between-groups analysis of variance procedures were conducted to investigate the main effects of a) question-type, b) interviewer demographic (race/gender), and c) interviewer/interviewee demographic alignment (race/gender). The interaction effects between question-type and interviewer demographic (race/gender), as well as between question-type and

interviewer/interviewee demographic alignment (race/gender) were also produced. The results are presented in Table 2-3.

Table 2-3

Comparison of the Difference (Outcome Variable: Authenticity)

Comparison	Source	df	F	Sig.	Partial Eta Squared
Question-Type / Race	Question-type	2	14.029	<.001	.174
	Interviewer Race	1	1.337	.250	.010
	Q* IR	2	2.217	.113	.032
Question-Type / Gender	Question-type	2	7.532	<.001	.102
	Interviewer Gender	1	.007	.934	.000
	Q*IG	2	.466	.628	.007
Question-Type / Race Alignment	Question-type	2	11.147	<.001	.144
	Race Alignment	1	1.373	.243	.010
	Q*RA	2	3.349	.038	.048
Question-Type / Gender Alignment	Question-type	2	11.646	<.001	.149
	Gender Alignment	1	.075	.785	.001
	Q*GA	2	.300	.741	.004

Note. N=139

As shown in Table 2-3, when the independent variables are interviewer race and question-type, results showed that there was a statistically significant main effect for question-type, $F(2, 133) = 14.03$, $p < .001$, and the effect size was large (partial eta squared = .174). Post-hoc comparisons using the Tukey HSD test indicated that the mean score for the group that responded to the Mentor question ($M = 69.45$, $SD = 28.42$) was significantly different from the group that responded to the Race question ($M = 91.08$, $SD = 12.90$) and the group that responded to the Success question ($M = 83.37$, $SD = 20.05$). The main effect for interviewer race, $F(1, 133) = 1.34$, $p = .25$, did not reach statistical significance. When the independent variables were question-type and interviewer gender, no statistically significant results occurred except there was main effect for question-type ($F[2, 133] = 7.53$, $p < .001$, eta = .102). Results showed that

interviewer and interviewee race/gender alignment had no main effect on the authenticity score of interviewee responses. However, the interaction effect between question-type and interviewer/interviewee race alignment was statistically significant, F (2, 133) = 3.35, p = .038.

Outcome: Emotional Tone

Using the interviewee's emotional tone score as outcome variable, an equivalent set of two-way between-groups analysis of variance procedures were conducted. The results obtained from the analysis are summarized in Table 2-4.

Table 2-4

Comparison of the Difference (Outcome variable: Emotional Tone)

Comparison	Source	df	F	Sig.	Partial Eta Squared
Question-Type / Race	Question-type	2	19.206	<.001	.224
	Interviewer Race	1	.040	.843	.000
	Q* IR	2	3.314	.039	.047
Question-Type / Gender	Question-type	2	18.733	<.001	.220
	Interviewer Gender	1	1.506	.222	.011
	Q*IG	2	4.434	.014	.063
Question-Type / Race Alignment	Question-type	2	18.862	<.001	.221
	Race Alignment	1	1.349	.247	.010
	Q*RA	2	.704	.496	.010
Question-Type / Gender Alignment	Question-type	2	18.467	<.001	.217
	Gender Alignment	1	.640	.425	.005
	Q*GA	2	.182	.834	.003

Note. N=139

As presented in Table 2-4, results for the outcome of emotional tone indicated that the interaction effect between question-type and interviewer race was statistically significant, F (2, 133) = 3.31, p =.04. There was a statistically significant main effect for question-type, F (2, 133) = 19.21, p < .001, and the effect size was large (partial eta squared =.224). Again, post-hoc comparisons using the Tukey HSD test indicated that the mean score for the group that

responded to the Mentor question (M = 53.68, SD = 21.49) was significantly different from the group that responded to the Race question (M = 39.79, SD = 18.06) and the group that responded to the Success question (M = 65.75, SD = 22.38). The main effect for interviewer race (F [1, 133] =.40, p =.84) did not reach statistical significance. When the independent variables were question-type and interviewer gender, the interaction effect between question-type and interviewer gender was statistically significant, F (2, 133) = 4.43, p = .01. There was also a large statistically significant main effect for question-type (F [2, 133] = 18.86, p < .001, eta = .221). Consistent with prior analyses, post-hoc comparisons using the Tukey HSD test indicated that the mean score for the group that responded to the Mentor question (M = 53.68, SD = 21.49) was significantly different from the group that responded to the Race question (M = 39.79, SD = 18.06) and the group that responded to the Success question (M = 65.75, SD = 22.38) . The main effect for interviewer gender, F (1, 133) = 1.51, p =.22 did not reach statistical significance. Lastly, Table 2-4 showed that interviewer and interviewee race/gender alignment had no main effect on the emotional tone score of interviewee responses.

When considering the conditions of question-type by interviewer (success; race; mentor), interviewer demographics (race/gender), interviewer and interviewee demographic alignment (race/gender) on the interviewer effect, neither interviewer race, interviewer gender nor the alignment of interviewer and interviewee demographic (race/gender) emerged as a factor significantly associated with the interviewee's authenticity or emotional tone score. However, the interaction between question-type and interviewer/interviewee race alignment did affect interviewee's performance during the interview, as there were statistically significant effects on the authenticity score. Additionally, the interaction between question-type and interviewer's race and gender affects significantly on emotional tone score.

Discussion

In the present study, we examined LIWC-22 derived authenticity and emotional tone scores computed from 57 interviewees' responses to assess the impacts of question-type, interviewer demographics (race/gender), and interviewer and interviewee demographics alignment (race/gender). This study represents the first attempt to implement this data analytic tool (i.e., LIWC-22) as part of an evaluation effort. Another motivation for this study was to assess the viability of using LIWC-22 to assist with formative evaluation, in this case, to evaluate the data collection process more systematically and efficiently.

In this study, we found large statistically significant main effects were limited only to question-type. Regardless of the demographics of the interviewer or the pairing of the interviewer and interviewee, certain interviewers produce interviewee responses with lower average authenticity and emotional tone scores. Given that sensitive topics involving race/ethnicity (e.g., Adida et al., 2016; Kim et al., 2018; van Bochove et al., 2015) are particularly prone to interviewer effects, the results of our analyses of main effects, including the absence of significant effects for race and gender, are somewhat surprising, though encouraging for ongoing data collection practices given the absence of significant effects for race and gender. The results suggest there may be opportunities to improve the phrasing of the specific interview questions (in particular, the mentor question) to promote rich data generation. However, ongoing monitoring with LIWC-22 to assess refinements of specific questions is needed. According to the results, the research team must also investigate which factors contribute to the variability in the outcomes seen among the interviewers. Given the role of the interviewer as an instrument for data collection in a qualitative study, the results suggest a closer review of the interview

transcripts from the interviewers identified as underperforming against procedural criteria or other possible quality standards. It is possible that further training may be warranted.

Interaction effects were also present between question-type and interviewer/interviewee race alignment for the outcome of authenticity as well as between question-type and both interviewer race and gender for emotional tone. The results support the theoretical suppositions of the "interviewer effect" concept that interviewers' race/ethnicity with respondent's characteristics is one of the top characteristics that affect response quality (West & Blom, 2017).

Contrary to our expectations, the alignment of interviewer and interviewee demographics (race/gender) had no main effect on the interviewees' performance in the interview as measured by the LIWC-22 summary variables. The results were inconsistent with the previous studies, which have shown gender and demographic matching are two of the top five characteristics that affect response quality (West & Blom, 2017); specifically, females tend to collect higher-quality data (Davis et al., 2010). Our findings suggest that in this particular qualitative research study context, the alignment of interviewer and interviewee demographics (race/gender) seems to have little influence on the quality of the data being collected.

With respect to the border RoE field, the community of evaluation consistently calls for more research on evaluation (Christie, 2003; Henry & Mark, 2003; Mark, 2008; Coryn et al., 2016). There is a need to identify effective and credible strategies and tools to facilitate evaluation to make the evaluation process efficient and less affected by evaluator bias (Coryn et al., 2017). Our research findings provide several important implications for evaluators in the monitoring of an interview data collection process with a focus on the "Input" and "Process" dimensions (Figure 2-1) of the CIPP model (Stufflebeam & Zhang, 2017). However, it is

reasonable to assume similar practices could be used to evaluate other qualitative data types or even other aspects of the evaluation process (e.g., context assessment and stakeholder selection).

Conclusions

Exploring the application of LIWC within our research reveals that it offers substantial benefits in analyzing qualitative data, particularly in identifying nuanced patterns and trends that might otherwise go unnoticed. Its ability to quantitatively analyze linguistic features provides a unique lens through which researchers can examine the emotional and cognitive aspects of verbal communication. This, in turn, has allowed for a deeper understanding of interviewer effects in our study, highlighting the significance of linguistic cues in shaping interview dynamics during data collection process.

Although we identified several interviewer effects in our evaluation effort, there were four distinct groups of participants at different stages of their doctoral journeys. This led us to wonder whether other influences exist, like the interviewee's stage in their doctoral program. The nature of interviewee responses may have been contextual, even when they were asked the same exact question, the interviewees may have interpreted the questions differently based on how long they had been connected to their program. Such a realization highlights how study design might also introduce uncertainty and inconsistency in the way the questions were phrased among different interviewers conducting interviews. In this regard, further studies aiming to refine this data science practice need to place more emphasis on introducing controls for the variability of the questions, interviewer, and the group of the interviewee to further support the investigation. However, because the focus is on identifying efficient real-world practices, which may not conform to uniformity in function, we still feel LIWC-22 offers a practical option to consider when reasonable uniformity is achieved.

The aim of the present study was not to call into question the validity of the existing interview process or the individual responses provided by the research participants. Instead, it was to test a strategy to efficiently identify areas of possible consideration that might narrow down where to concentrate complementary qualitative analyses to inform whether the interviewer effects are impactful on the broader project. In this way, the use of LIWC-22 was envisioned as a formative process. As such, examining a host of factors is not a reasonable practice for promoting efficiency in evaluation, so evaluators must prioritize their aims.

The focus of the study is on using data analytic strategies to explore the feasibility of using LIWC to efficiently assess the transcribed data for possible interviewer effects to support a formative evaluation. One consequence of this data analytics method, particularly in this context, is that it symbolically reduces the compilation of powerful and often consequential lived experiences shared by each student to just a few data points. In doing so, our intention is most certainly not to minimize the value of the stories provided by the interviewees but instead focus on the role of the evaluator to ensure that these lived experiences are captured in an upholding manner. There is absolutely no substitute for in-depth qualitative data analysis to examine these stories and learn from them.

Limitations

In addition to the challenges discussed above, there are other challenges and limitations to highlight. The first limitation was this was an exploratory study. Our aim was to use LIWC-22 to efficiently target the part of the data collection that needs to be helped with. Because of this we did not account for multiplicity across all analyses. Additionally, the sample sizes were small as we only had a total of 57 interviews to use for this study. Furthermore, the data derived from the original study was not collected for the purpose of evaluation, and therefore, as noted above,

lacked consistency as it was generated through a semi-structured protocol. More consistency is needed to improve confidence in the interviewer effects findings. First, the data must be organized and easy to work with. As noted in the methods section, data cleaning processes sometimes involve our own judgments about the intention of responses, and it is possible we misinterpreted the intent of responses. To improve this aim, data collection should be streamlined. Evaluators may even plan to equate responses using consistent questions for a more sophisticated calibration across individuals. Additionally, several summary variables appeared to align quite well with project aims, particularly authenticity and emotional tone, but this may not be the case in future studies. Even in the present study, the use of each summary variable is most certainly an over-simplification of quality, therefore caution is urged in the interpretation of findings for each outcome. Evaluators will need to develop and refine other functional constructs beyond authenticity and emotional tone that can be applied to a variety of studies.

Lessons Learned

In addressing the challenges identified through our evaluation, we emphasize the need for carefully managing the interview process including how interviewers ask questions, who conducts the interviews, and who we interview to get reliable and useful information. Utilizing LIWC-22 as a tool becomes beneficial when we can ensure that participants interpret the questions consistently. Our aim was to make sure we collect our data in a clear and consistent way. This was particularly important given the exploratory nature of our study and the limited number of interviews we worked with. Moreover, we advocate for expanding our analytical framework beyond basic measures like authenticity and emotional tone, urging a deeper exploration into the data. This strategy is designed to refine our analysis, ensuring it is both

comprehensive and accessible, and facilitates a more nuanced understanding of our findings, striking a balance between detailed inquiry and overarching insights.

Implications

The present research also provides great implications for further study. For example, this study focused on the Input and Process dimensions in the CIPP model; other dimensions of the CIPP model definitely can be a good angle to consider when evaluating a research project. LIWC-22 dictionary has more than 120 built-in variables, and there is the possibility to conduct a factor analysis on the qualitative results to determine further which factor is the most accurate indicator for useful interviews. Moreover, since our study is context-sensitive, if we want to apply it to other contexts, there is the possibility to create your own dictionary.

Interviewer effects are an important topic, particularly in the current context of monitoring and evaluating the qualitative data collection process. Corresponding to the definition and continuing need for more RoE, identifying credible tools to assist evaluators in the field is an irresistible trend. In this study, we illustrate how identifying interviewer effects using LIWC-22 can enhance the design, implementation, and monitoring of a qualitative data collection research project.

Funding

This work is supported by the National Science Foundation under Grant Agreements #1920753 and #2207795. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Adida, C. L., Ferree, K. E., Posner, D. N., & Robinson, A. L. (2016). Who's asking? Interviewer coethnicity effects in African survey data. *Comparative Political Studies*, *49*(12), 1630–1660. https://doi.org/10.1177/0010414016633487

Alkin, M. C., & Vo, A. (2018). *Evaluation essentials: From A to Z*. Guilford Press.

Al-Shanawani, H. M. (2019). Evaluation of self-learning curriculum for kindergarten using Stufflebeam's CIPP model. *SAGE Open*, *9*(1), 215824401882238. https://doi.org/10.1177/2158244018822380

American Mathematical Society. (2018). Mathematical and Statistical Sciences Annual Survey. Retrieved from http://www.ams.org/profession/data/annual-survey/2018Survey-NewDoctorates-Report.pdf

Benstead, L. J. (2013). Effects of interviewer–respondent gender interaction on attitudes toward women and politics: Findings from Morocco. *International Journal of Public Opinion Research*, *26*(3), 369–383. https://doi.org/10.1093/ijpor/edt024

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: A quantitative analysis. *Assessing Writing*, *12*(3), 213–227. https://doi.org/10.1016/j.asw.2008.02.001

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice–theory relationship: New directions for evaluation* (Vol. 97, pp.7–36). San Francisco, CA: Jossey-Bass.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological

change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. https://doi.org/10.1111/j.0956-7976.2004.00741.x

Coryn, C. L. S., Ozeki, S., Wilson, L. N., Greenman, II, G. D., Schro¨ter, D. C., Hobson, K. A., . . . Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation*, *37*(2), 159–173. https://doi.org/10.1177/1098214015611245

Coryn, C. L. S., Wilson, L. N., Westine, C. D., Hobson, K. A., Ozeki, S., Fiekowsky, E. L., … Schröter, D. C. (2017). A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation*, *38*(3), 329–347. https://doi.org/10.1177/1098214016688556

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171– 192. http://dx.doi.org/10125/44329

Cutler, A. D., Carden, S. W., Dorough, H. L., & Holtzman, N. S. (2020). Inferring grandiose narcissism from text: LIWC versus machine learning. *Journal of Language and Social Psychology*, *40*(2), 260–276. https://doi.org/10.1177/0261927x20936309

Davidson, R. J., & Kaszniak, A. W. (2015). Conceptual and methodological issues in research on mindfulness and meditation. *American Psychologist*, *70*(7), 581-592. https://doi.org/10.1037/a0039512

Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell., & K. Resnicow (2010). Interviewer effects in public health surveys. *Health Education Research*, *25*(1), 14–26. https://doi.org/10.1093/her/cyp046

del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science*, *40*(6), 749–760. https://doi.org/10.1177/0165551514547842

Dexter, L. A. (1970). *Elite and specialized interviewing*. Evanston, IL: Northwestern University Press.

Esgaiar, E., & Foster, S. (2019). Implementation of CIPP model for quality evaluation at Zawia university. *International Journal of Applied Linguistics and English Literature*, *8*(5), 106. https://doi.org/10.7575/aiac.ijalel.v.8n.5p.106

Foster, J. (1994). The dynamics of gender in ethnographic research: A personal view. In R. G. Burgess (Ed.), *Studies in qualitative methodology 4: Issues in qualitative research*. Greenwich, CT: JAI Press.

Francis, M. E., & Pennebaker, J. W. (1992). Putting stress into words: The impact of writing on physiological, absentee, and self-reported emotional well-being measures. *American Journal of Health Promotion*, *6*(4), 280–287. https://doi.org/10.4278/0890-1171-6.4.280

Fournier, D. M. (1994). [Review of *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs*, by The Joint Committee on Standards for Educational Evaluation]. *Journal of Educational Measurement*, *31*(4), 363–367. http://www.jstor.org/stable/1435400

Fuchs, M. (2009). Gender-of-interviewer effects in a video-enhanced web survey. *Social Psychology*, *40*(1), 37–42. https://doi.org/10.1027/1864-9335.40.1.37

Geng, S., Niu, B., Feng, Y., & Huang, M. (2020). Understanding the focal points and sentiment of learners in MOOC reviews: A machine learning and SC-LIWC-based approach.

British Journal of Educational Technology, *51*(5), 1785–1803. https://doi.org/10.1111/bjet.12999

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.*), The practice–theory relationship in evaluation: New directions for evaluation* (Vol. 97, pp. 69–80). San Francisco, CA: Jossey-Bass.

Humphreys, A., & Wang, R. J.-H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104

Imahori, E. (2018). Linguistic expressions of depressogenic schemata. *Working Papers in Applied Linguistics & TESOL*, *18*(2), 20–32. http://tesolal.columbia.edu/

Kim, N., Krosnick, J. A., & Lelkes, Y. (2018). Race of interviewer effects in telephone surveys preceding the 2008 U.S. presidential election. *International Journal of Public Opinion Research*, *31*(2), 220–242. https://doi.org/10.1093/ijpor/edy005

Kühne, S. (2020). Interpersonal perceptions and interviewer effects in face-to-face surveys. *Sociological Methods & Research*, *52*(1), 299-334. https://doi.org/10.1177/0049124120926215

Langley, A., & Meziani, N. (2020). Making interviews meaningful. *The Journal of Applied Behavioral Science*, *56*(3), 370–391. https://doi.org/10.1177/0021886320937818

Leone, T., Sochas, L., & Coast, E. (2021). Depends who's asking: Interviewer effects in demographic and health surveys abortion data. *Demography*, *58*(1), 31–50. https://doi.org/10.1215/00703370-8937468

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

Loosveldt, G. (2008). Face-to-face interviews. In de, L. E. D., Hox, J. J., & Dillman, D. A (Eds.), *International Handbook of Survey Methodology* (pp. 201-220). Routledge. https://doi.org/10.4324/9780203843123.ch11

Mark, M. M. (2008). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111–134). New York, NY: Guilford.

Martin, J. D. (2020). It depends on who's asking: Interviewer gender effects on credibility ratings of male and female journalists in six Arab countries. *International Journal of Public Opinion Research*, *33*(1), 18–37. https://doi.org/10.1093/ijpor/edz053

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation*. Jossey-Bass, a Wiley Brand.

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: Examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, *27*(5-6), 655–669. https://doi.org/10.1080/10494820.2019.1610453

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Okahana, H., Klein, C., Allum, J., & Sowell, R. (2018). STEM doctoral completion of underrepresented minority students: Challenges and opportunities for improving participation in the doctoral workforce. *Innovative Higher Education*, *43*(4), 237–255. https://doi.org/10.1007/s10755-018-9425-3

Oliver, K. M., Houchins, J. K., Moore, R. L., & Wang, C. (2020). Informing Makerspace outcomes through a linguistic analysis of written and video-recorded project assessments.

International Journal of Science and Mathematics Education, *19*(2), 333-354. https://doi.org/10.1007/s10763-020-10060-2

Patton, M. Q. (2015). *Qualitative evaluation and research methods* (4th ed.). Thousand Oaks, CA: Sage.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, *10*(6), 601–626. https://doi.org/10.1080/026999396380079

Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*(3), 271–282. https://doi.org/10.1006/jrpe.2002.2349

Scriven, M. (1982). *The logic of evaluation*. Edgepress.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. Jossey-Bass.

Seidman, I. (2013). *Interviewing as qualitative research* (4th ed.). New York: Teachers College Press.

Skolits, G. J., Morrow, J. A., & Burr, E. M. (2009). Reconceptualizing evaluator roles. *American Journal of Evaluation*, *30*(3), 275–295. https://doi.org/10.1177/1098214009338872

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications*. Jossey-Bass.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. John Wiley & Sons.

Stufflebeam, D. L., & Zhang, G. (2017). *The CIPP evaluation model: How to evaluate for improvement and accountability*. The Guilford Press.

van Bochove, M., Burgers, J., Geurts, A., de Koster, W., & van der Waal, J. (2015). Questioning ethnic identity: Interviewer effects in research about immigrants' self-definition and feelings of belonging. *Journal of Cross-Cultural Psychology*, *46*(5), 652–666. https://doi.org/10.1177/0022022115576961

Volkov, B. B. (2011). Beyond being an evaluator: The multiplicity of roles of the internal evaluator. *New Directions for Evaluation*, *2011*(132), 25–42. https://doi.org/10.1002/ev.394

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

Wuyts, C., & Loosveldt, G. (2020). Measurement of interviewer workload within the survey and an exploration of workload effects on interviewers' field efforts and performance. *Journal of Official Statistics*, *36*(3), 561–588. https://doi.org/10.2478/jos-2020-0029

Yang, M.-L., & Yu, R.-R. (2008). The interviewer effect when there is an education gap with the respondent: Evidence from a survey on Biotechnology in Taiwan. *Social Science Research*, *37*(4), 1321–1331. https://doi.org/10.1016/j.ssresearch.2008.05.008

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2010). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Williams, J., Shea, C., & Misullis, K. (2011). Using the context, input, process, and product evaluation model (CIPP) as a comprehensive framework to guide the planning, implementation, and assessment of

service-learning programs. *Journal of Higher Education Outreach and Engagement*, *15*(4), 57-84. https://openjournals.libs.uga.edu/jheoe/article/view/901

CHAPTER 3 [ARTICLE III]: PROJECT MONITORING AND EVALUATION: APPLICATION OF DATA ANALYTICS FOR INTERVIEW RESPONSE GRADING

Zhi Li *Carl Westine*

For the development of scientific research, outstanding research projects are needed, and external evaluation is recognized as a respected strategy to improve grant research studies. The external evaluator's job is typically defined to judge the quality of research practices and products based on a set of defined criteria and standards (Scriven, 1983; Fournier, 1994), but during implementation, the purpose of the evaluation often is formatively defined with a goal of project improvement. As such, formative evaluation tends to be focused more on program processes (though not necessarily). It frequently involves a more consultative or collaborative approach, given the underlying need to implement changes for improvement (Tsipianitis & Mandellos, 2022) Activities vary across projects but frequently include gathering and interpreting administrator/staff and participants' feedback on their experiences, tracking various implementation forms with multiple measures to identify best practices, assessing the need for and use of project resources, and secondary analysis of project impact data (Tsipianitis & Roumelioti, 2021). Additional practices are undoubtedly possible. In this study, we used data analytic strategies to inform the formative evaluation of a qualitative research project. Project data was used to identify the underlying linguistic sources of variation based on ratings of data richness as defined by the project research team. This study was expected to show a new method for evaluators to leverage tools to perform formative evaluations efficiently.

Loud and Mayne (2014) suggest that organizations are keen on endorsing activities that have the potential to enhance their value; these are echoed in the Research on Evaluation (RoE) literature, which states that evaluators seek new tools and strategies to accomplish their tasks

efficiently. For example, Azzam and Harman (2016) examined the use of Amazon's Mechanical Turk (MTurk) website for crowd-sourcing to rate and code interview transcripts, noting consistency and value in the process. Additional tools within data visualization have also been shown to enhance other evaluation-related activities, such as the analysis of clustered data (Trochim & McLinden, 2017) and to promote better evaluation reporting and interpretation of evaluation findings (Evergreen & Metzner, 2013). They also can play an important supplemental role, such as validating program theory and strengthening the reliability of existing qualitative analyses (Harman & Azzam, 2018).

Linguistic Inquiry and Word Count (LIWC) software is a text analysis tool designed to quantitatively assess the linguistic features of both spoken and written materials, as noted by Boyd et al. (2022). LIWC includes a 'dictionary' that facilitates the quantitative evaluation of qualitative data by categorizing words into predefined measures according to the characteristics of their content and calculating the frequency of words corresponding to each measure (Boyd et al., 2022). In its most recent version, LIWC-22 outputs variables that include linguistic variables such as drive, cognition, affect, social processes, culture, lifestyle, physical, states, motives, perception, conversation, and all punctuation (Boyd et al., 2022, Table 3, p.15) based on daily conversation transcription. These variables, as well as several multidimensional summary variables, have been applied to analyze thousands of spoken and written text forms, giving credence to their applied use in the social sciences. Therefore, it is important to examine its value in the field of evaluation.

Formative evaluation often prioritizes internal values as practical performance measures for convenience and underscores the importance of process use (Patton, 2007) as a mechanism to promote implementation fidelity and further clarifies program practices. Evaluation scholars

have repeatedly shown that proper evaluation use is tied to engaging intended uses for intended users (King & Alkin, 2018; Patton, 2008). In most formative settings, particularly for a formative evaluation of a grant-funded project, the intended users are the project's principal investigators and those working on the front lines of implementation where possible changes will occur.

For a qualitative research project, research team leads and interviewers function as key stakeholders in managing and carrying out the project's research mission and incorporating formative evaluation findings. Thus, for this study, which draws upon a qualitative research project to document the lived experiences of mathematics doctoral students from underrepresented groups, we utilized the research team members' perspectives of value with respect to the interview data and examined its factor structure using the full scope of LIWC variables. In this manner, this study showed how evaluators can systematically and diagnostically use LIWC to empirically assess the project team's internal values and formatively evaluate the project team's data collection practices.

LIWC

LIWC has been used throughout the social sciences to extract meaning from subtle speech and written text patterns. In the field of Psychology, LIWC has been used to explore the relationship between the psychological process and word usage in daily conversation (e.g., Cutler et al., 2020; Rude et al., 2004; Zhao et al., 2016). LIWC has been utilized in the field of Computer Science as a tool for natural language processing aimed at extracting computable features from text data available online (e.g., del Pilar Salas-Zárate et al., 2014; Golbeck et al., 2011; Sell & Farreras, 2017). Its use has also spread to various fields and subfields in Education (e.g., Robinson et al., 2013; Moore et al., 2019; Yoo & Kim, 2014), Linguistics (e.g., Carroll, 2007; Imahori, 2018), and Cognition (Pennebaker & Francis,1996).

One reason for its widespread use is that LIWC dimensions have demonstrated good reliability and validity evidence through years of work and research. For example, Pennebaker and Francis (1996) required expert judges to rate students' essays for 12 LIWC categories, scaling from 1 to 7. Results showed medium to high correlations between human ratings and LIWC variables in the dimension of emotion and cognitive process. Similar studies involving expert ratings have been conducted, including Alpers et al. (2005). Zhao et al. (2016) also note that comparing human ratings and LIWC variables contributes to LIWC validity. Some purely psychometric studies compile multiple forms of reliability and validity information, including correlations that demonstrate good reliability of dimensions across writing samples, comparisons of factor structures of LIWC dimensions across various forms of text noting high levels of consistency, and correlations that support criterion-related validity through predictive, concurrent, and convergent comparisons (e.g., Pennebaker & King, 1999). Tausczik and Pennebaker (2010) summarized predictive validity research from hundreds of studies that use LIWC dimensions. Recent studies continued the trend of generating predictive validity evidence using LIWC across multiple settings and contexts, including academic performance (Lewine et al., 2019; Robinson et al., 2013; Yoo & Kim, 2014), personality (Mairesse et al., 2007), mental health status (Hao et al., 2013). This plethora of reliability and validity evidence supports the continued use of LIWC variables and their applications; however, Boyd et al. (2022) note the fundamental importance of predictive validity in expanding the use of LIWC for social research.

Applications of LIWC in social science research vary in their format and use of one or more LIWC variables. Some research employs only one category of LIWC (Moore et al., 2019), and some select particular LIWC variables that fit their research purposes (Lewine et al., 2019). Some use all variables of categories of word use in LIWC to reveal the factor structure of the

LIWC variables (predictor) that are highly correlated with their outcome variable (Robinson et al., 2013).

LIWC has been applied in various fields but has not been strategically embedded into formal evaluations. However, Pennebaker and Francis's (1996) correlations of human rating with LIWC variables and Robinson et al. (2013) correlation, factor analysis, and linear regression provide insight for evaluators to adapt LIWC for practice, particularly in a formative setting. When an evaluation is formative, evaluators may use exploratory techniques to gain insight into the program processes. Robinson et al. (2013) used a three-step process to examine the utility of LIWC for predicting final course performance from students' written self-introductions. The researchers first conducted a bivariate correlation of 80 LIWC-15 variables and final course performance. Secondly, they ran a principal component analysis with varimax rotation to reduce the 20 predictor variables correlated with final course performance. Lastly, they enter the scores for the eight identified factors into a multiple linear regression analysis to determine which factors predict final course performance when controlling for other remaining factors.

Robinson et al. (2013) provide valuable insight on using LIWC for prediction with factor analysis. Such a strategy aligns well with evaluation scenarios where qualitative inquiry is the evaluand or qualitative data is a central component of decision-making (e.g., context assessment, stakeholder selection, standard setting, interpreting findings). Using a qualitative phenomenology grant-funded research project, we explored the practicability of applying data analytics to identify subtle, hidden patterns based on linguistic features in interview transcripts that can be used to improve future data collection practices. In this way, we tested the use of LIWC to formatively evaluate data collection practices and devise practical suggestions for enhancing the ability of the research team to collect the type of data they desire.

Phenomenology & Rich Data Collection

Central to phenomenological research is the endeavor to grasp the essence of human experiences. In practice, the phenomenological study involves in-depth interviews, observations, or other data collection methods to gather rich narratives from participants. In a phenomenological study, rich data emerges from the comprehensive accounts of participants' lived experiences. Such data can situate the participants' emotions, perceptions, beliefs, recollections, and interpretations. By aiming for richness in data, researchers seek to obtain a comprehensive understanding of the phenomenon under study, embracing its complexity and the multifarious elements that constitute individual experiences.

For interview data, qualitative researchers generally intend to capture the depth, detail, and nuanced understandings of certain phenomena when participants describe their personal experiences. For interviews to fulfill their potential as a tool for gaining insight into a person's life and experiences, the data produced must be sufficiently detailed and rich, enabling the research team to employ them for comprehensive analysis and thick description (Brekhus et al., 2005; Ponterotto, 2006).

According to Schultze and Avital (2011), the richness of the data can be interpreted in two distinct ways: a) focusing on the data provided by each individual, richness describes the deeply-nuanced descriptions of events, and b) richness refers to the overall value of the dataset and its generative qualities to produce a diversity of new ideas and insights. Collectively, the richness of the data (both types) enables the researcher to produce transferable knowledge and meaning from the analysis by promoting thick descriptions in reporting (Merriam & Tisdell, 2016). For monitoring and evaluation purposes, we worked off the assumption that when individual data is collected in such a way that it is rich (both types), it will be more beneficial to

the research team and valuable in supporting the accomplishment of the project objectives (i.e., identifying themes, and developing counter stories.)

This study has two primary objectives, which seek to explore one particular use of LIWC within the evaluation process. The first is to better understand the construct of the human judgments of the qualitative interview data. Understanding the principal aspects underlying the construct of rich data, as defined by the research team, will better understand what the research team values in the collected qualitative data and what is linguistically driving this perceived value. The second is to demonstrate the potential of using LIWC to aid evaluators in conducting formative evaluations. Demonstrating the predictive ability of LIWC would suggest that evaluators could use the findings to offer practical advice for the research team on collecting desired data. For example, the research team could retrain interviewers or modify data collection materials and methods to ensure the collected data reflects what is desired. Additionally, this process should give rise to avenues for developing tailored automatic indices for application in other real-world evaluation processes involving valuing. The following research questions guided the study.

1. Which LIWC variables predict project team members' ratings of the richness of interview data?
2. Using LIWC output variables, what is the factor structure of team members' ratings of the richness of interview data?
3. What factors are most closely associated with higher ratings of interviews, and how much variability in the ratings do they explain?

Method

Dataset

The dataset comes from a National Science Foundation research study documenting successful doctoral students' experiences, perspectives, and stories and recent PhDs from historically marginalized racial groups in mathematics, including Black students, Latinx students, and Indigenous students. The grant project research team collected 57 in-depth interviews. All semi-structured interviews followed a general interview protocol aligned to four key topical areas: background, environment, mentoring, and academics. Each interview was recorded and transcribed and lasted between 60 and 200 minutes. After transcription, the interview transcripts were scrubbed for identifying information and assigned a random identifier.

Sample

For the present study, the sample dataset includes interview transcripts from 57 participants on three consistently asked questions. The questions are as follows: (1) What does success in graduate school mean to you? (2) How do you think race has played a role in your math experience? (3) What is the ideal mentoring relationship looks like to you? Across all the transcripts reviewed, information relevant to the question of success was found in 51 transcripts, to the question of race in 46 transcripts, and to the question of mentorship in 42 transcripts. There are a total of 139 responses. Within 57 interviewees concerning race/ethnicity, the majority of the interviewees identified as Black/African American, comprising 54.39% of the sample with a count of 31 individuals. Latino/a interviewees represented 33.33%, amounting to 19 individuals, while those identifying as Multiracial formed the smallest group at 12.28%, totaling seven individuals (Table 3-1). Regarding gender distribution, the sample was relatively balanced. Females accounted for a slight majority, making up 52.63% of the interviewees,

corresponding to 30 individuals. Males represented 47.37% of the sample, with a total of 27 individuals (Table 3-2).

Table 3-1

Distribution of Interviewee's Race/Ethnicity in the Sample

Race/Ethnicity	N	Percentage
Black/African American	31	54.39%
Latino/a	19	33.33%
Multiracial	7	12.28%

Table 3-2

Distribution of Interviewee's Gender in the Sample

Gender	N	Percentage
Male	27	47.37%
Female	30	52.63%

In order to conduct data analysis, 139 responses from the sampled transcripts were first rated by the research team and then analyzed using LIWC-22. There are four raters, each contributing to the assessment of responses with varying quantities. Each response was rated by three raters. Raters were rotated during this rating process. Rater 1, a White male, rated the highest number of responses, totaling 113. He was closely followed by Rater 2, a White female, who rated 109 responses. Rater 3, a Black/African American female, contributed ratings for 91 responses, while Rater 4, a Black/African American male, rated 103 responses (Table 3-3). The composition of raters in terms of race/ethnicity and gender was balanced, with an equal representation of two White and two Black/African American raters and an equal gender distribution of two males and two females.

Table 3-3

Raters' Information

Rater	Race/Ethnicity	Gender	# of Rated Responses
1	White	Male	113
2	White	Female	109
3	Black/African American	Female	91
4	Black/African American	Male	103

Analysis Procedure

The analysis was largely motivated by Robinson et al.'s (2013) three-step procedure for data analysis and factor analysis to predict the research team's rating score of the interview transcripts. However, we first define the rating by asking the research team to rate each interview response based on data richness with thick descriptions on the transcripts. We further performed an exploratory factor analysis (EFA). Principal axis factoring (PAF) with oblimin rotation was used instead of using principal component analysis (PCA) compared with Robinson et al. (2013).

To answer the research questions, we first had four research team members rate the interview responses' richness with thick description using a 1 to 5 scale (1—not very rich; 2—somewhat rich; 3—rich; 4—very rich; 5—extremely rich). The higher the score, the richer the interview, as perceived by the research team.

To promote a high standard of reliability and validity in the rating process, the raters engaged in a series of calibration exercises prior to the formal evaluation of responses. Initially, three preparatory meetings were convened to establish a consensus on the rating criteria and to clarify any uncertainties regarding the rating standards. During these sessions, raters discussed and refined the criteria, ensuring a shared understanding of the evaluation process. Subsequently, each rater independently rated a set of five randomly selected responses. These preliminary ratings served as a benchmark for consistency across raters. Following this initial rating exercise,

the raters reconvened to compare their evaluations, discuss any discrepancies, and reconcile differences in their interpretations of the rating scale. This collaborative approach was designed to align the raters' perspectives and cultivate a uniform rating standard. By discussing their ratings and reaching a consensus, the raters were able to calibrate their criteria, thereby enhancing the reliability of the ratings. This iterative process of independent rating followed by collective discussion was instrumental in promoting both the reliability and validity of the subsequent ratings.

The analysis of the research team's ratings revealed a distribution of ranges that offers insight into the level of agreement among raters (Table 3-4). Specifically, the lowest range (0), indicating unanimous agreement, accounted for 12.95% of the ratings, suggesting that in these instances, raters were in complete concordance regarding the value of the data. The most common range observed was 1, comprising 51.08% of the ratings, which signifies a high degree of agreement among raters, albeit with slight variations in their evaluations. A range of 2, reflecting moderate agreement, was noted in 25.90% of cases, while a range of 3, indicative of more substantial differences in rating perspectives, constituted 9.35% of the total. Notably, the highest range (4), which implies the greatest divergence in rater evaluations, was rare, occurring in only 0.72% of instances. Altogether, the distribution of ranges underscores the predominantly high level of consensus among the research team's raters, with the majority of ratings falling within a narrow range, thereby reflecting a robust alignment in their assessments of the data's value. We also calculate Kendall's Tau-b to show the correlations between the raters. The highest is 0.512, and the lowest is 0.222, with an average of 0.398, which suggests strong consistency among raters in their assessments.

Table 3-4

Distribution of Ranges from Research Team's Ratings

Range	Count	Percentage
0	18	12.95%
1	71	51.08%
2	36	25.90%
3	13	9.35%
4	1	0.72%
Total	139	

To answer the first research question, we conducted a correlation analysis on the 117 LIWC-22 variables and the research team's median rating of interviewees' responses to see how many LIWC-22 outcome variables significantly correlate with the research team's rating score. Spearman's rank-order test was used to identify significant correlations, which indicated that the variable should be retained for further analysis.

To answer the second research question, Principal axis factoring (PAF) with oblimin rotation was performed on the LIWC-22 variables significantly correlated with the research team's median rating.

The final research question is based on factor analysis. We conducted a multiple linear regression analysis on the factor scores to check which ones predict the research team's rating score when controlling for other factors. The adjusted R-squared value indicated the total variance in the outcome explained by the set of significant predictors.

Results

Bivariate Correlation of the LIWC Variables and Raters' Rating

A bivariate correlation analysis was conducted on the 117 LIWC variables and the research team's median rating of interviewees' responses. Of the 117 potential predictor

variables, 15 of them yielded a significant correlation with final course performance (see Table 3-5).

Table 3-5

Predictor Variables Correlated with Raters' Rating

Variable	Description/Most frequently used exemplars	R
Words per sentence	Average words per sentence	-.180*
Personal pronouns	I, you, my, me	-.170*
3rd person singular	he, she, her, his	.262**
Impersonal pronouns	that, it, this, what	.215*
Prepositions	to, of, in, for	-.180*
Conjunctions	and, but, so, as	-.174*
Cognitive processes	but, not, if, or, know	-.176*
Insight	know, how, think, feel	-.175*
Discrepancy	would, can, want, could	-.303**
Certitude	really, actually, of course, real	-.205*
Negative tone	bad, wrong, too much, hate	.218**
Anger	hate, mad, angry, frustr	.229**
Interpersonal conflict	fight, kill, killed, attack	.285**
Female references	she, her, girl, woman	.267**
Curiosity	scien*, look* for, research*, wonder	.177*

Note. N = 139.
Description/Most frequently used exemplars taken from Boyd et al. (2022)
*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Exploratory Factor Analysis

A Principal Axis Factoring (PAF) analysis with oblimin rotation was employed to discern the underlying structure of predictor variables that correlate with raters' ratings using SPSS 28.0.0.0. After initially screening 15 items, we retained 13 items based on scree plot, eigenvalues greater than 1, and loading greater than .30. Next, we performed the oblimin method of rotation, allowing the factors to correlate because research based on LIWC variables are related (Boyd et

al., 2022). The analysis elucidated three distinct factors, which accounted for a cumulative variance of 44.025%. The factors were labeled based on the theoretical interpretation of the variable loadings (Table 3-6).

Factor 1, termed "Refined/Reflective Storytelling," emerged as the most significant factor, boasting an eigenvalue of 5.139 and accounting for 31.755 of the variance. This factor was primarily defined by high loadings of conjunctions (0.709) and prepositions (0.608) and negative loadings of impersonal pronouns (-0.711) and curiosity (-0.705), suggesting a narrative style that is elaborate and introspective yet less inquisitive and impersonal.

Factor 2, designated as "Certain/Confident Language," presented with an eigenvalue of 1.838, explaining an additional 8.553% of the variance. This factor was notably influenced by high loadings on certitude (0.795) and discrepancy (0.650), indicating an assertive and definitive language style often utilized in the context of articulating contrasts or disagreements.

Factor 3, identified as "Contextualized Relationships/Conflicts," had an eigenvalue of 1.205 and contributed 3.717% to the explained variance. It was most strongly characterized using third-person singular pronouns (0.722), which may signal a narrative focus on third-party individuals. Additional moderate loadings on interpersonal conflict (0.354), negative tone (0.307), and female references (0.304) suggest that this factor encapsulates language relating to social dynamics, possibly with a nuanced focus on gendered discourse or contentious interactions.

The factor correlation matrix revealed that Factor 1 and Factor 3 were inversely correlated ($r = -0.45$), indicating that narratives characterized by "Refined/Reflective Storytelling" tend to diverge from those associated with "Contextualized Relationships/Conflicts." Conversely, the correlation between Factor 1 and Factor 2 was

relatively small ($r = -0.176$), suggesting a slight tendency for refined and reflective narratives to be less characterized by "Certain/Confident Language." The correlation between Factor 2 and Factor 3 was negligible ($r = 0.017$), implying that assertive language and the contextualization of relationships or conflicts operate almost independently within the dataset.

These inter-factor relationships underscore the multidimensional nature of language as it pertains to raters' evaluations. The identified factors capture distinct but interconnected communication elements, offering a nuanced understanding of the linguistic dimensions that shape narrative perception.

Table 3-6

Factor Loadings of Predictor Variables Correlated with Raters' Rating

Variables	Factors		
	1	2	3
Impersonal pronouns	-.711		
Conjunctions	.709		
Curiosity	-.705	.336	
Prepositions	.608		
Cognitive processes	.603	-.433	
Words per sentence	.568		
Insight	.469	-.445	
Certitude		.795	
Discrepancy		.650	
Anger			
3rd person singular			.722
Interpersonal conflict			.354
Negative tone			.307
Female references			.304
Personal pronouns			
Eigenvalues	5.139	1.838	1.205
Percent of variance explained	31.755	8.553	3.717
Factor correlations			
1	———	———	———
2	0.176	———	———
3	0.454	0.017	

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

Multiple Regression Analysis with the Factor Scores as Predictors of Raters' Rating

A multiple regression analysis was conducted to investigate the predictive validity of the three factors derived from the Principal Axis Factoring (PAF) analysis on raters' ratings. The model summary indicates that the collective predictors explain 17.6% of the variance in raters' ratings (Table 3-7). The regression equation was statistically significant, $F(3, 135) = 9.637$, $p < .001$, indicating that the model significantly predicts the outcome variable.

The beta coefficients suggest the strength and direction of the relationship between each predictor and the outcome variable. The "Refined/Reflective Storytelling" factor had a significant positive association with raters' ratings ($p = .012$), suggesting that as the storytelling becomes more refined and reflective, raters' ratings increase. On the other hand, "Certain/Confident Language" was not a significant predictor ($p = .651$), indicating that this type of language does not significantly influence raters' ratings.

The third factor, "Contextualized Relationships/Conflicts," had the strongest positive association with raters' ratings ($p < .001$), suggesting that narratives that effectively contextualize relationships and conflicts are associated with higher ratings by raters.

The regression coefficients provide evidence that while certain/confident language does not significantly contribute to raters' evaluations, refined/reflective storytelling and contextualized relationships/conflicts play essential roles, with the latter being the more robust predictor.

These results underscore the importance of the narrative context and the manner of storytelling in influencing evaluative outcomes. They provide empirical support for including such factors in the assessment of narrative quality and effectiveness, especially in contexts where raters' evaluations are pivotal.

Table 3-7

Multiple Regression Analysis Summary for the Factor Score as Predictors of Raters' Rating

Predictor Variables	B	Std. Error	Beta	t	Sig.
(Constant)	3.03	.158		19.140	<.001
Refined/reflective storytelling	.003	.001	.289	2.555	.012
Certain /confident language	.007	.016	.052	.453	.651
Contextualized relationships/conflicts	.477	.096	.421	4.965	<.001

Note. $R^2 = .176$

Discussion

This study's exploration into the feasibility and value of employing LIWC-22 pre-established variables to predict raters' ratings of data richness marks a significant advancement in project monitoring and evaluation within an NSF grant-funded qualitative research. Integrating LIWC-22 into the formative evaluation process has demonstrated its potential to efficiently assist evaluators in identifying high-quality data defined by the research team. This aligns with Loud and Mayne's (2014) emphasis on the necessity for organizations to adopt value-adding activities, further supported by Boyd et al. (2022), who detailed LIWC's capabilities for automatic text analysis. Such analysis not only contributes to the practical evaluation of linguistic features that enrich interview transcripts but also underscores the software's utility in enhancing the quality of data collection methodologies.

The application of an exploratory factor analysis (EFA), largely adapted from Robinson et al. (2013), has unveiled distinct linguistic patterns that resonate with raters' perceptions of data richness. Factors like "Refined/Reflective Storytelling" and "Contextualized Relationships/Conflicts" emerged as significant predictors of higher ratings. These findings echo the reliability and validity of LIWC variables demonstrated in earlier studies by Pennebaker and Francis (1996), among others, reinforcing the effectiveness of LIWC-22 in identifying nuanced linguistic elements that signify valuable data.

Beyond data analysis, LIWC-22's contribution extends to offering actionable insights for research teams to refine data collection protocols. This practical application not only addresses the study's objective of improving interview data quality but also illustrates the broader implications of utilizing LIWC-22 in formative evaluation efforts. It is also possible to note the idea of using it as a tool to understand and further calibrate teams' thinking about valuing. It’s

clear that the typical calibration process didn't really work out well for reliability, but perhaps the identification of terms and clarification about the factors that are seen as contributing to richness could be used formatively to build consensus around richness and promote a more informed discussion. For example, this could conceivably be useful for concepts other than richness that require valuing during the evaluation process, such as determining performance metrics or standards. Such enhancements are invaluable for ensuring that research practices and outcomes align with the high standards of quality and richness desired in qualitative research defined by the research team.

The integration of additional insights from the document segments into our discussion has further enriched our understanding of LIWC-22's role in evaluating and enhancing qualitative research. By operationalizing data valuation through identifying and categorizing specific linguistic features, LIWC-22 bridges the intuitive and empirical assessment of data richness. This enables research teams to articulate and quantify the value of data collected, transforming the qualitative data collection process into a more targeted and effective endeavor.

Moreover, LIWC-22's capacity to provide examples of words and phrases associated with these valuable linguistic patterns offers practical guidance for interviewers. For example, "he," "she," "but," "also," and "as" are part of the linguistic repertoire associated with valued data richness, enabling interviewers to subtly guide interviews in directions that are likely to generate the desired rich, thick descriptions. This practical application of LIWC-22's findings empowers research teams to enhance the overall quality of data collection by fostering an interview environment conducive to "Refined/Reflective Storytelling" and the sharing of "Contextualized Relationships/Conflicts."

Conclusions

In conclusion, we have delved into the realm of qualitative research monitoring and evaluation through the lens of Linguistic Inquiry and Word Count (LIWC-22), revealing its potential to significantly enhance the assessment and understanding of qualitative research. However, there were some limitations. The original study's semi-structured design was not inherently focused on evaluation, leading to variability in interview questioning and, consequently, in the data collected. This variability, alongside a concentrated examination of responses to three specific questions, "Success, Race, and Mentorship," resulted in a limited sample size, impacting our findings' broader applicability and strength. Additionally, the agreement among the raters responsible for evaluating the richness of the interview responses presents an opportunity for further calibration and enhancement of the evaluation process. This could influence the evaluation of linguistic characteristics pinpointed by LIWC-22 as indicators of richness.

Despite these constraints, our exploration illuminates the significant role of LIWC-22 as a complement to traditional qualitative analysis methods. Acknowledging the inherent complexity of qualitative research studies and the nuanced process of qualitative coding is important. This study does not aim to challenge or replace traditional qualitative coding methods but rather to explore the potential of LIWC-22 as a supplementary tool in the qualitative data analysis process. Given the time-consuming nature of analyzing qualitative data, we sought to determine whether LIWC-22 could offer additional insights to research teams. The ease and speed of LIWC-22 present it as an attractive supplementary resource that could potentially enhance the analysis of qualitative interview data.

This study not only underscores the utility of data analytics in qualitative evaluation but also sets the stage for future research aimed at integrating such tools more comprehensively. It emphasizes the need for future investigations to address the constraints encountered, expand the analytical scope, and refine the use of automated text analysis in qualitative research evaluation.

Looking ahead, the implications of our work suggest that using LIWC as part of the qualitative analysis process is valuable and necessitates further exploration across various research contexts. The adaptability of LIWC, particularly the potential for creating custom dictionaries tailored to specific client requirements or research purposes, opens new avenues for making qualitative analysis more nuanced and contextually relevant. This flexibility enhances the tool's applicability, allowing researchers to capture the information they require in the most meaningful ways to their specific studies.

Incorporating LIWC into future qualitative research endeavors offers an exciting opportunity to validate its effectiveness further and explore its utility across different domains and research objectives. The strategic application of linguistic analysis, augmented by the capability to customize the LIWC dictionary, can significantly contribute to methodological rigor, ethical considerations, and the overall quality of qualitative research evaluation. Consequently, this study not only contributes to the literature on applying data analytics in evaluation but also lays a foundation for future research that seeks to refine and expand the use of automated text analysis tools like LIWC-22, thereby enhancing the depth, accuracy, and insight of qualitative research methodologies.

Funding

This work is supported by the National Science Foundation under Grant Agreements #1920753 and #2207795. Any opinions, findings, and conclusions or recommendations

expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Barr Taylor, C. (2005). Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior*, *21*(2), 361–376. https://doi.org/10.1016/j.chb.2004.02.008

Azzam, T., & Harman, E. (2016). Crowdsourcing for quantifying transcripts: An exploratory study. *Evaluation and Program Planning*, *54*, 63–73. https://doi.org/10.1016/j.evalprogplan.2015.09.002

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *18*(6), 1–13.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app

Brekhus, W. H., Galliher, J. F., & Gubrium, J. F. (2005). The need for thin description. *Qualitative Inquiry*, *11*(6), 861–879. https://doi.org/10.1177/1077800405280663

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: A quantitative analysis. *Assessing Writing*, *12*(3), 213–227. https://doi.org/10.1016/j.asw.2008.02.001

Costello, A. B., & Osborne, O. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(7), 1–9. https://doi.org/https://doi.org/10.7275/jyj1-4868

Cutler, A. D., Carden, S. W., Dorough, H. L., & Holtzman, N. S. (2020). Inferring grandiose narcissism from text: LIWC versus machine learning. *Journal of Language and Social Psychology*, *40*(2), 260–276. https://doi.org/10.1177/0261927x20936309

del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science*, *40*(6), 749–760. https://doi.org/10.1177/0165551514547842

Evergreen, S., & Metzner, C. (2013). Design principles for data visualization in evaluation. *New Directions for Evaluation*, *2013*(140), 5–20. https://doi.org/10.1002/ev.20071

Fournier, D. M. (1994). The program evaluation standards: How to assess evaluations of educational programs. *Journal of Educational Measurement, 31*(4), 363-367.

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*. https://doi.org/10.1145/1979742.1979614

Hao, B., Li, L., Li, A., & Zhu, T. (2013). Predicting mental health status on social media. *Cross-Cultural Design. Cultural Differences in Everyday Life*, 101-110. https://doi.org/10.1007/978-3-642-39137-8_12

Harman, E., & Azzam, T. (2018). Towards program theory validation: Crowdsourcing the qualitative analysis of participant experiences. *Evaluation and Program Planning*, *66*, 183–194. https://doi.org/10.1016/j.evalprogplan.2017.08.008

Imahori, E. (2018). Linguistic expressions of depressogenic schemata. *Working Papers in Applied Linguistics & TESOL*, *18*(2), 20–32.

King, J. A., & Alkin, M. C. (2018). The centrality of use: Theories of evaluation use and influence and thoughts on the first 50 years of use research. *American Journal of Evaluation*, *40*(3), 431–458. https://doi.org/10.1177/1098214018796328

Lehman, A. (2005). *Jmp for basic univariate and multivariate statistics: A step-by-step guide*. SAS Institute.

Lewine, R., Warnecke, A., Davis, D., Sommers, A., Manley, K., & Calebs, B. (2019). Gender and affect: Linguistic predictors of successful academic performance among economically disadvantaged first year college students. *International Journal for the Scholarship of Teaching and Learning*, *13*(1). https://doi.org/10.20429/ijsotl.2019.130102

Loud Marlène Läubli, & Mayne, J. (2014). *Enhancing evaluation use: Insights from Internal Evaluation Units*. Sage.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–500. https://doi.org/10.1613/jair.2349

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: a guide to design and implementation* (Fourth edition.). Jossey-Bass.

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: Examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, *27*(5-6), 655–669. https://doi.org/10.1080/10494820.2019.1610453

Patton, M. Q. (2007). Process use as a usefulism. *New Directions for Evaluation*, *2007*(116), 99–112. https://doi.org/10.1002/ev.246

Patton, M. Q. (2008). *Utilization-focused Evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, *10*(6), 601–626. https://doi.org/10.1080/026999396380079

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Ponterotto, J. G. (2006). Brief note on the origins, evolution, and meaning of the qualitative research concept thick description. *Qualitative Report*, *11*(3), 538–549.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions. *Journal of Language and Social Psychology*, *32*(4), 469–479. https://doi.org/10.1177/0261927x13476869

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research. *Information and Organization*, *21*(1), 1–16. https://doi.org/10.1016/j.infoandorg.2010.11.001

Scriven, M. (1993). *Hard-won lessons in program evaluation*. Jossey-Bass.

Sell, J., & Farreras, I. G. (2017). LIWC-ing at a century of introductory college textbooks: Have the sentiments changed? *Procedia Computer Science*, *118*, 108–112. https://doi.org/10.1016/j.procs.2017.11.151

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927x09351676

Trochim, W. M., & McLinden, D. (2017). Introduction to a special issue on concept mapping. *Evaluation and Program Planning*, *60*, 166–175. https://doi.org/10.1016/j.evalprogplan.2016.10.006

Tsipianitis, D., & Mandellos, G. (2022). The value of formative evaluation in an education program. *International Journal of Applied Systemic Studies*, *9*(4), 381. https://doi.org/10.1504/ijass.2022.126770

Tsipianitis, D., & Roumelioti, I. (2021). Formative evaluation for intelligence quality management in an education program. case study. *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. https://doi.org/10.1109/iisa52424.2021.9555576

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3). https://doi.org/10.33151/ajp.8.3.93

Yoo, J., & Kim, J. (2013). Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education*, *24*(1), 8–32. https://doi.org/10.1007/s40593-013-0010-8

Zhao, N., Jiao, D., Bai, S., & Zhu, T. (2016). Evaluating the validity of Simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services. *PLOS ONE*, *11*(6). https://doi.org/10.1371/journal.pone.0157947

OVERALL CONCLUSION

The series of studies presented in this dissertation collectively work to bridge the longstanding theory-practice gap in the field of research on evaluation (RoE). The collective findings from this dissertation significantly advance our understanding of evaluation context and tool utilization within the RoE. By meticulously examining three distinct but interconnected studies, this work not only echoes but also builds upon the foundational calls by esteemed scholars for a deeper, more nuanced exploration of RoE through a) enhancing the RoE field by expanding upon prior categorization and knowledge extraction efforts from empirical studies and b) advancing the practice by pinpointing innovative tools aimed at elevating the efficiency and efficacy of evaluation practice. Each article contributes uniquely towards this aim by blending theoretical insights with practical applications, thereby enhancing the effectiveness and efficiency of evaluation practices.

Summary of Findings

The First Article

The first study, "The Role of Context: A Synthesis of Empirical Research on Evaluation Context," delves into the intricate dynamics of evaluation context and its influence on evaluation practices. Through a meticulous qualitative analysis of recent RoE literature, it uncovers how the evaluation context dimensions—encompassing evaluator characteristics, stakeholder perspectives, organizational/program features, and historical/political landscapes—play a pivotal role in affecting evaluation practices and shaping evaluation outcomes. The study expands on Coryn et al. (2017) to systematically categorize and analyze empirical findings related to these context dimensions (Vo, 2013). It identifies critical relationships between context descriptors and

their impact on evaluation, highlighting the nuanced manner in which context elements interplay to affect evaluation practice.

The Second Article

The second study, "Using Data Analytics to Monitor and Evaluate Qualitative Data Collection Processes for Interviewer Effects," explores the effectiveness of the innovative Linguistic Inquiry and Word Count (LIWC) software in identifying and mitigating interviewer effects within the context of qualitative research data collection. By analyzing interview transcripts with LIWC, the research uncovers distinct linguistic patterns that are influenced by the dynamics between interviewers and interviewees, including variations in authenticity and emotional tone based on the interviewer's race/gender and the alignment of interviewer/interviewee demographics. Significantly, it finds that the type of question posed by interviewers is associated with both the authenticity and emotional tone of the responses. Moreover, the study highlights the interaction effects, demonstrating that the alignment between interviewer and interviewee demographics in terms of race and gender, coupled with the nature of questions asked, significantly affects the perceived authenticity and emotional tone of interview responses. However, the alignment of interviewer and interviewee demographics alone does not show a main effect on these variables, suggesting the complexity of factors that contribute to interviewer effects.

The Third Article

The third study, "Project Monitoring and Evaluation: Application of Data Analytics for Interview Response Grading," builds on the advancements made in enhancing data collection through LIWC, as detailed in the second study, itlooks deeper into data analytics further to refine

our understanding and application of interview response grading. It showcases the next steps in our exploratory journey into formative evaluation methodologies.

This study utilizes exploratory factor analysis and multiple regression analysis to identify linguistic patterns that correspond with the research team's perceptions of data richness in interview responses. Key findings reveal that linguistic features categorized under "Refined/Reflective Storytelling" and "Contextualized Relationships/Conflicts" are significant predictors of higher evaluation ratings. These insights underscore LIWC's potential as a valuable tool for formative evaluation, offering a novel approach to improving interview methodologies and enhancing the overall quality of qualitative research efforts.

Connecting to the Evaluation Process

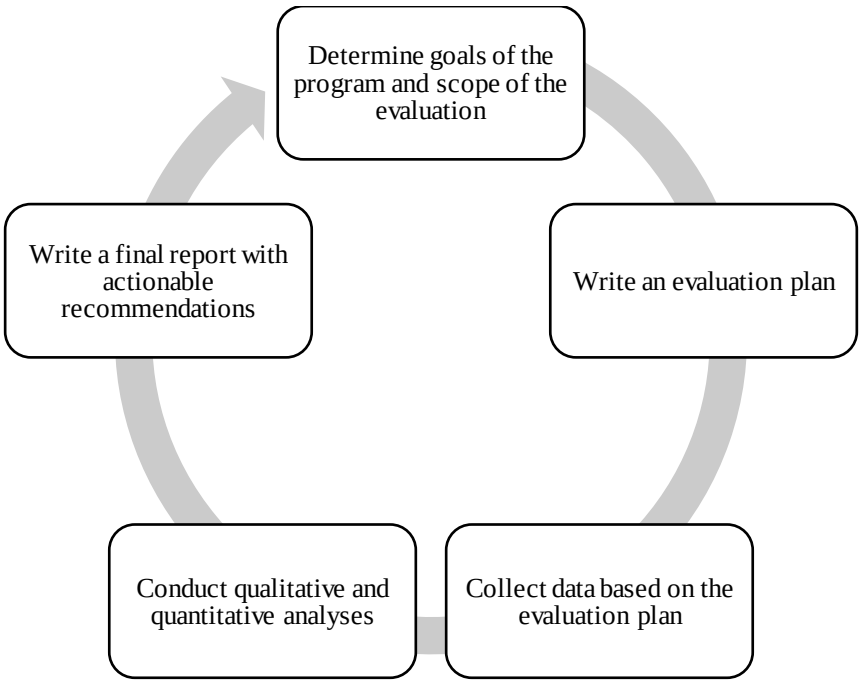

Figure 4-1. Typical Program Evaluation Process (Source: https://www.dallasisd.org/Page/42560)

In all, this dissertation explores how RoE can inform and improve evaluation practice. The first study sets the stage by emphasizing the importance of understanding the context within which an evaluation takes place, a step that is both foundational and critical to the process. This

initial inquiry not only echoes RoE's call for deeper methodological refinement and knowledge extraction from empirical research but also aligns perfectly with the first phase of the Typical Program Evaluation Process depicted in Figure 4-1. It lays the groundwork for a relevant and targeted evaluation.

The journey continues with the second and third articles, which shift focus to the innovative data analytic capabilities of the LIWC tool. Exploring the application of this sophisticated instrument responds to the RoE's invitation to promote the evaluation field forward by introducing tools to improve the efficiency and effectiveness of evaluation practices. These two studies correspond with the data collection and analysis phases, the third and fourth phases of the Typical Program Evaluation Process illustrated in Figure 4-1. Through practical application in monitoring and evaluating qualitative research projects, LIWC demonstrates its utility, illustrating the potential for such tools to revolutionize program evaluation.

Implications for RoE

The insights garnered from the first article have significant implications for the RoE field. First, they underscore the necessity of a holistic and systematic approach to understanding evaluation contexts. Evaluators are encouraged to consider not only the program and organizational contexts but also the broader socio-political and historical environments in which evaluations occur. This comprehensive perspective can guide evaluators in designing and implementing more effective and responsive evaluation strategies.

Second, the findings advocate for the development of flexible, context-sensitive evaluation frameworks. Such frameworks can assist evaluators in navigating the complex features of the evaluation context and enable them to tailor their approaches to meet the unique needs and challenges of each evaluation scenario.

Lastly, the first article sets a precedent for future RoE studies to further investigate the multifaceted nature of the evaluation context. By continuing to explore and explain the complex interconnections between context dimensions and evaluation practice, RoE can contribute to the advancement of a more nuanced, informed, and effective evaluation methodology.

Building on the foundation laid by Article 1, which highlighted the importance of understanding the context of evaluations, Articles 2 and 3 offer significant advancements for the RoE field through the practical application of the LIWC tool in a specific qualitative research study context.

Article 2 introduces LIWC as a means to enhance the analysis of qualitative interview data, specifically addressing the interviewer effect—a step forward in the RoE community's ongoing efforts to ensure data reliability and validity. This tool provides evaluators with the means to delve deeper into the narrative quality, allowing for a more detailed and nuanced examination of interview content. Article 3 takes this a step further by demonstrating how LIWC can predict the richness of interview data, thus aiding evaluators in making informed decisions about data collection and analysis. This predictive ability means that recommendations for improving data collection can be based on solid evidence, leading to more meaningful and comprehensive qualitative data.

Together, the insights from these two articles suggest that incorporating technology like LIWC can revolutionize the way evaluators work. This approach supports RoE's objectives to render the evaluation process more empirical, accurate, and objective. It also resonates with the need for flexible tools adaptable to various contexts, reaffirming the importance of context-aware evaluation frameworks, as discussed in Article 1.

Looking ahead, this body of work underscores the potential for further exploration into the integration of technology within RoE. By continuing to test and implement innovative tools across different evaluation settings, the field is set to evolve towards practices that are not only data-rich and methodologically solid but also deeply rooted in the context of each evaluation, aligning with the practical needs of programs and stakeholders. The collective insights from these articles guide RoE towards an integrative future where technology enhances the clarity and effectiveness of evaluations.

Implications for Evaluation Practice

The findings from the second article provide crucial guidance for evaluating qualitative research, particularly in projects involving interview data. These findings highlight the importance of acknowledging and mitigating potential interviewer effects. Enhancements in training programs could include strategies to minimize these effects through careful question formulation and an awareness of how interviewer characteristics might influence responses. Additionally, the application of the LIWC underscores the benefits of using automated text analysis to identify these effects, suggesting broader applications for such tools in evaluating qualitative research. By creating tailored dictionaries, evaluators can adapt LIWC to suit specific evaluation needs, thereby enhancing data collection protocols. These improved protocols can facilitate the matching of interviewers to interviewees based on demographic characteristics to enhance response authenticity and emotional tone.

The third article builds on these insights by illustrating the significant role of LIWC in formative evaluations. Incorporating LIWC enables a deeper understanding of interview data richness, guiding the development of more effective data collection methodologies. This approach not only helps in identifying specific linguistic patterns that signify data richness but

also aids in creating systematic frameworks for evaluating interview data. Such frameworks enhance the objectivity and accuracy of assessments, improving the reliability and validity of the evaluations. Moreover, the insights gained from LIWC usage can refine training programs, enabling researchers and evaluators to elicit more meaningful responses from participants.

Furthermore, the integration of data analytics tools like LIWC into the evaluation process represents a transformative advance in qualitative research methodologies. This integration facilitates more accurate, efficient, and influential evaluations, setting new standards for quality. It also points towards the potential for incorporating more advanced AI technologies in future evaluations, which could further enhance the sophistication of data analysis and the overall evaluation process, as suggested by Nielsen (2023).

By combining the implications from both articles, it becomes clear that the strategic use of tools like LIWC in the early stages of project involvement can significantly improve the effectiveness and efficiency of data management in qualitative research evaluations. This approach not only enhances the reliability and authenticity of collected data but also fosters the development of innovative evaluation methodologies that could reshape the landscape of qualitative research.

Future Directions

Building on the foundational insights from various articles, several strategic directions exist for advancing our understanding and application of evaluation methodologies.

Next Step for Research

Inspired by the first article, the next step involves developing a new framework or reporting standard that systematically incorporates context dimensions into evaluation processes. This framework should integrate the latest RoE studies up to 2024, reflecting contemporary

research and practices. Additionally, there is a valuable opportunity to expand the investigation into different RoE domains identified by Coryn et al. (2017), particularly focusing on evaluation activities. Exploring how various RoE domains influence evaluation processes and outcomes could provide a more comprehensive understanding of evaluation practices, thereby enriching the field of RoE with nuanced insights and practical tools.

Stemming from insights in the second article on the use of the LIWC tool, the next steps include enhancing training for interviewers who consistently yield lower scores in Authenticity and Emotional Tone. Such training should not only focus on improving interviewing skills but also on increasing awareness of how interviewer behaviors influence these metrics. There is also a call for greater standardization in semi-structured interview protocols to ensure consistency across interviews. Expanding the research to incorporate other factors from the CIPP (Context, Input, Process, Product) model could provide a more thorough understanding of the evaluation process. Furthermore, adapting LIWC variables more closely aligned with specific research goals and applying LIWC in other research contexts or designing custom dictionaries tailored to specific evaluative needs can broaden the applicability of LIWC in qualitative research evaluations.

Based on the third article, future research directions include broadening the study's sample size to encompass a wider array of interview questions beyond the three consistently asked questions (success, race, mentorship). This would provide a richer dataset for analysis and deeper insights into linguistic patterns across different thematic areas. Similar to the second article, further standardization in collecting and analyzing interview data is necessary. Establishing more uniform procedures for conducting interviews and applying LIWC analysis will enhance the reliability and validity of the findings. Additionally, testing this approach in

other research contexts would validate the utility of LIWC in diverse settings and potentially innovate new methodologies for qualitative data analysis.

Collectively, these steps will significantly enhance the robustness and relevance of tools like LIWC in improving the quality and effectiveness of evaluations that involve qualitative data, thereby paving the way for innovative evaluation methodologies that cater to a diverse range of research needs.

Advancing Data Analytics in Evaluation Practice

This dissertation explores LIWC's feasibility as an innovative tool in qualitative research formative evaluation, setting the stage for a comprehensive discussion of its revolutionary impact, benefits, limitations, and potential to integrate data analytics and, to a greater extent, AI in future evaluation practices (Montrosse-Moorhead, 2023). By systematically categorizing words into psychologically meaningful categories, LIWC facilitates a nuanced understanding of the underlying themes, emotions, and cognitive processes present in qualitative data. This capability complements traditional qualitative analysis methods by offering evaluators and researchers an additional lens through which to examine data, enhancing efficiency and contributing greater depth and breadth to the analysis.

Building on this foundation, it is crucial to note the specific advantages that LIWC offers, shedding light on how it enriches the evaluative process. LIWC processes large volumes of text data rapidly, providing immediate insights into the linguistic patterns embedded in qualitative data. This speed is invaluable in evaluative settings where time and resource constraints are a consideration. Furthermore, LIWC enables evaluators to efficiently focus more intentionally on supporting the earlier stages of qualitative research, especially on providing formative feedback on data collection processes and contributing to interpreting the results, rather than only

replicating analyses through manual coding. However, this efficiency should enhance analyses by ensuring that relevant linguistic patterns are not overlooked due to human error. This objectivity can help mitigate ethical concerns related to interpretation bias or implicit bias and, therefore, offers a complementary role to traditional qualitative aims. Finally, by providing a standardized method for text analysis, LIWC can and should be routinely and rigorously tested for cost-effectiveness in its use across other evaluative actions involving the use of qualitative data (e.g., context assessment, stakeholder selection, valuing, setting standards, planning for use, even meta-evaluation). Although presently evaluation-specific dictionaries do not exist, resources such as the evaluation thesaurus (Scriven, 1993) have long been available. Additionally, efforts to understand and measure evaluation capacity (e.g., Nielsen et al., 2011; Taylor-Ritzler et al., 2013) may offer insight to advance this development effort.

Having explored the benefits, it is equally important to consider the boundaries within which LIWC operates, highlighting its limitations in the context of qualitative research formative evaluation. Despite its strengths, LIWC is not designed to replace traditional qualitative analysis methods. It lacks the ability to fully grasp the nuanced meanings, contexts, and subtleties that human analysis can uncover. Therefore, LIWC should be viewed as a complementary tool that enhances rather than supplants the rich, detailed insights gained through conventional qualitative research methods.

With these insights into LIWC's advantages and limitations, we now turn our attention to its promising future avenues, particularly how it may contribute to a new era of utilizing AI in evaluation practices. The successful integration of LIWC in formative evaluation hints at a promising future where AI plays a significant role in RoE. AI technologies, with their ability to learn and adapt, could further refine the analysis of qualitative data, offering even more

sophisticated insights into written and verbal data. Future AI tools could potentially interpret nuances and contexts with greater accuracy, bridging the gap between quantitative rigor and qualitative depth.

Moreover, the evolution of AI in evaluation is anticipated to make qualitative research more accessible by automating complex analyses and making it easier for evaluators to uncover hidden patterns and insights within their data. As AI technology continues to advance, the potential for its application in formative evaluation expands, promising a future where evaluators are equipped with an even broader arsenal of tools to enhance the accuracy, efficiency, and depth of their analyses (Reid, 2023).

Considering these insights, this dissertation draws upon Christie's (2003) call for more RoE, particularly leveraging the work of Henry and Mark (2003), Mark (2008), Vo and Christie (2015), and Coryn et al. (2017). These studies collectively underscore the significant impact of context on evaluation, providing a robust empirical foundation for this exploration. It progresses to investigate the application of LIWC in formative evaluations of qualitative research, marking a notable advancement in the RoE field. This exploration highlights the substantial promise of incorporating sophisticated analytical tools in evaluation practices.

References

Alkin, M. C. (2003). Evaluation theory and practice: Insights and new directions. In C. A. Christie (Ed.), *The practice–theory relationship: New directions for evaluation* (Vol. 97, pp. 89–91). San Francisco, CA: Jossey-Bass.

Azzam, T., & Harman, E. (2016). Crowdsourcing for quantifying transcripts: An exploratory study. *Evaluation and Program Planning*, *54*, 63–73. https://doi.org/10.1016/j.evalprogplan.2015.09.002

Azzam, T., & Robinson, D. (2013). GIS in evaluation: Utilizing the power of Geographic Information Systems to represent evaluation data. *American Journal of Evaluation, 34*(2), 207–224. https://doi.org/10.1177/1098214012461710

Bernstein, I. N., & Freeman, H. E. (1975). *Academic and entrepreneurial research*. New York: Russell Sage.

Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on Cross-Cultural Evaluation. *American Journal of Evaluation*, *30*(4), 457–494. https://doi.org/10.1177/1098214009349865

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice–theory relationship: New directions for evaluation* (Vol. 97, pp.7–36). San Francisco, CA: Jossey-Bass.

Coryn, C. L. S., Ozeki, S., Wilson, L. N., Greenman, II, G. D., Schro¨ter, D. C., Hobson, K. A., . . . Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation, 37*, 159–173.

Coryn, C. L. S., Wilson, L. N., Westine, C. D., Hobson, K. A., Ozeki, S., Fiekowsky, E. L., … Schröter, D. C. (2017). A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation, 38*(3), 329–347. doi: 10.1177/1098214016688556

Coryn, C. L. S., & Westine, C. D. (Eds.). (2015). *Contemporary trends in evaluation research* (Vols. I-IV). (Sage benchmarks in social research methods). London, England: Sage.

Cousins, J.B., Goh, S.C., Clark, S., & Lee, L.E. (2004). Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge base. *Canadian Journal of Program Evaluation, 19,* 99-141.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56(3),* 331-364.

Dallas Independent School District (n.d.). *Program evaluation process.* Evaluation and Assessment / Program Evaluation Process. https://www.dallasisd.org/Page/42560

Evergreen, S., & Metzner, C. (2013). Design principles for data visualization in evaluation. *New Directions for Evaluation*, *2013*(140), 5–20. https://doi.org/10.1002/ev.20071

Gedutis, A., Teresa Biagetti, M., & Ma, L. (2022). The challenges for research evaluation ethics in the social sciences. *Handbook on Research Assessment in the Social Sciences*. https://doi.org/10.4337/9781800372559.00032

Hansen, M., Alkin, M., & Wallace, T. (2013). Depicting the logic of three evaluation theories. *Evaluation and Program Planning*, *38*(C), 34–43. https://doi.org/10.1016/j.evalprogplan.2012.03.012

Harman, E., & Azzam, T. (2018). Towards program theory validation: Crowdsourcing the qualitative analysis of participant experiences. *Evaluation and Program Planning*, *66*, 183–194. https://doi.org/10.1016/j.evalprogplan.2017.08.008

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.*), The practice–theory relationship in evaluation: New directions for evaluation* (Vol. 97, pp. 69–80). San Francisco, CA: Jossey-Bass.

Montrosse-Moorhead, B. (2023). Evaluation criteria for artificial intelligence. *New Directions for Evaluation*, *2023*(178–179), 123–134. https://doi.org/10.1002/ev.20566

Morris, M., & Cohn, R. (1993). Program evaluators and ethical challenges: A national survey. *Evaluation Review, 17,* 621–642.

Nielsen, S. B. (2023). Disrupting evaluation? emerging technologies and their implications for the Evaluation Industry. *New Directions for Evaluation*, *2023*(178–179), 47–57. https://doi.org/10.1002/ev.20558

Nielsen, S. B., Lemire, S., & Skov, M. (2011). Measuring evaluation capacity—results and implications of a Danish study. *American Journal of Evaluation*, *32*(3), 324–344. https://doi.org/10.1177/1098214010396075

Johnson, K, Greenseid, *L.* O., Toal, S. A., King, J. A., Lawrenz, R, & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation,* 30(3), 377-410.

Leviton, L. C., & Hughes, E. F. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review,* 5(4), 525-548.

Mark, M. M. (2008). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111–134). New York, NY: Guilford.

Milzow, K., Reinhardt, A., Söderberg, S., & Zinöcker, K. (2019). Understanding the use and usability of research evaluation studies. *Research Evaluation, 28*(1), 94-. https://doi.org/10.1093/reseval/rvy040

O'Connor, T. (2023). Procedural and participatory ethics: Community-based evaluation in practice. *Evaluation Journal of Australasia*, *23*(2), 91–100. https://doi.org/10.1177/1035719x231166206

Patton, M. Q., Grimes, P. S., Guthrie, K. M., Brennan, N. J., French, B. D., & Blyth, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in public policy making* (pp. 141–184). Lexington, MA: Lexington Books.

Reid, A. M. (2023). Vision for an equitable AI world: The role of evaluation and evaluators to incite change. *New Directions for Evaluation*, *2023*(178–179), 111–121. https://doi.org/10.1002/ev.20559

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions. *Journal of Language and Social Psychology, 32*(4), 469–479. https://doi.org/10.1177/0261927x13476869

Scriven, M. (1993). *Hard-won lessons in program evaluation*. Jossey-Bass.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice, 14,* 237–242.

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5,* 19–25.

Szanyi, M., Azzam, T., & Galen, M. (2013). Research on evaluation: A needs assessment. *Canadian Journal of Program Evaluation, 27*, 39–64.

Taylor-Ritzler, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D. B., & Balcazar, F. E. (2013). Understanding and measuring evaluation capacity. *American Journal of Evaluation*, *34*(2), 190–206. https://doi.org/10.1177/1098214012471421

Trochim, W. M., & McLinden, D. (2017). Introduction to a special issue on concept mapping. *Evaluation and Program Planning*, *60,* 166–175. https://doi.org/10.1016/j.evalprogplan.2016.10.006

Vallin, L. M., Philippoff, J., Pierce, S., & Brandon, P. R. (2015). Research-on-evaluation articles published in the *American Journal of Evaluation*, 1998–2014. In P. R. Brandon (Ed.), *Research on evaluation*. *New directions for evaluation* (Vol. 148, pp. 7–15). San Francisco, CA: Jossey-Bass.

Vo, A. T. (2013). Visualizing context through theory deconstruction: A content analysis of three bodies of evaluation theory literature. *Evaluation and Program Planning, 38,* 44-52. doi:10.1016/j.evalprogplan.2012.03.013

Vo, A. T., & Christie, C. A. (2015). Advancing research on evaluation through the study of context. *New Directions for Evaluation, 148,* 43-55.

Webb, A. L., Schumacker, R. E., & Tilford, A. (2017). Synthesis of published articles from studies in educational evaluation, 2010-2015. *Journal of Education and Human Development*, *6*(1), 78–81. doi: 10.15640/jehd.v6n1a7

Weiss, C. H. (Ed.). (1977). *Using social research in public policy making*. Lexington, MA: Lexington Books.